DOCTOR OF PHILOSOPHY

Management, visualisation & mining of quantitative proteomics data

Ahmad, Yasmeen

Award date:
2012

DOCTOR OF PHILOSOPHY

Management, visualisation & mining of quantitative proteomics data

Yasmeen Ahmad

2012

University of Dundee

Management, Visualisation & Mining of Quantitative Proteomics Data

Yasmeen Ahmad BSc (Hons)

Doctor of Philosophy

University of Dundee
Nethergate, Dundee, DD1 4HN

March 2012

Table of Contents

LIST OF FIGURES **7**

LIST OF TABLES **11**

ACKNOWLEDGEMENTS **13**

CANDIDATE DECLARATION **15**

EXECUTIVE SUMMARY **17**

LIST OF PUBLICATIONS **19**

TALKS AND POSTER PRESENTATIONS **21**

ABBREVIATIONS **25**

INTRODUCTION **29**

CHAPTER 1: LITERATURE REVIEW **33**

1.1 Cell Biology & Proteomics 33

1.1.1 Genomics & Proteomics *33*

1.1.2 Mass Spectrometry *37*

1.1.3 Stable Isotope Labelling using Amino Acids in Cell Culture *40*

1.1.4 Challenges in the Proteomics Field *41*

1.2 Data Management Technology 43

1.2.2 Current Data Management in Life Sciences *43*

1.2.3 Comparison of Existing Software *49*

1.2.3 Experimental Data Workflow *51*

1.2.4 Workflow Management Systems *52*

1.2.5 Potential for Collective Data Analysis *53*

1.2.6 Data Recording & Collection *54*

1.2.7 Data Dissemination *55*

1.2.8 Public Data Repositories *57*

1.2.9 Data Standards *58*

1.3 Data Visualisation Approaches 60

1.3.1 Origins of Data Visualisation *60*

1.3.2 Types of Data Visualisations *62*

1.3.3 Potential for Data Visualisation in Life Sciences *63*

1.3.4 Problems with Data Visualisations *65*

1.3.5 Web-Based Visualisation Technologies *68*

1.4 Software Approaches for Biological Data Analysis 69

1.4.1 Super-Experiment Data Analysis *69*

1.4.2 Applying Business Intelligence for Super-Experiment Analysis *70*

1.4.3 Relational versus Multi-Dimensional Databases *72*

CHAPTER 2: METHODOLOGY 75

2.1 Software Development Approach 75

2.1.1 Developer Environment *75*

2.1.2 Ethnographic Observation *75*

2.1.3 New Ideas and Inspiration *76*

2.1.4 Domain Knowledge *76*

2.1.5 Understanding Users *76*

2.1.6 Researcher Expectations *77*

2.1.7 Function and Form *77*

2.1.8 Leadership *78*

CHAPTER 3: NUCLEOLAR PROTEOMICS DATABASE 79

3.1 Summary 79

3.2 Background 79

3.3 Nucleolar Proteome Database v 3.0 80

3.4 Technical Implementation 81

3.4.1 NOPdb3.0 Databases *82*

3.4.2 Application Programming Interface *83*

3.4.3 Application Security *84*

3.4.4 Client Side Interfaces *84*

3.5 Discussion 86

CHAPTER 4: PEPTRACKER - A TOOL FOR PROTEOMICS DATA MANAGEMENT & ANALYSIS 89

4.1 Summary 89

4.2 MsTrack – Laboratory Information Management System 90

4.2.1 Tag Cloud *94*

4.3 DataVault – Storage, Visualisation & Exploration of Quantitative Proteomics Data 95

4.3.1 Data Storage *96*
4.3.2 Browser View for Dataset Management *98*
4.3.3 Data View for Data Visualisation and Exploration *100*
4.4 ProteinLibrary – Protein Search and Specialised Protein Group Management *105*
4.4.1 Protein Search *105*
4.4.2 Protein Group Definition and Enrichment Analysis *107*
4.5 PepTracker Desktop Client 108
4.6 Technical Implementation 109
4.6.1 System Architecture *109*
4.6.2 Database Development *109*
4.6.3 Server-Side Setup *111*
4.6.4 PepTracker Task Scheduler *112*
4.6.5 Proteomics Server *113*
4.6.6 LDAP Authentication *113*
4.6.7 Graphical User Interfaces *113*
4.6.8 Protein Identification Strategies *114*
4.6.9 Protein Definitions *117*
4.7 Data Security & Quality Control 118
4.8 Discussion 119

CHAPTER 5: MULTIDIMENSIONAL ANALYSIS WITH IP EXPERIMENTS 125

5.1 Summary 125
5.2 Background 126
5.3 Results 131
5.3.1 Discriminating Specific from Non-Specific Interaction Partners *131*
5.3.2 Sun Model and the Protein Frequency Library *134*
5.3.3 Filtering of Protein Frequency Library using Experimental Parameters *137*
5.3.4 Application of PFL to Analysis of Multi-Protein Complexes *139*
5.3.5 Use of PFL to Normalise Datasets *141*
5.3.6 Comparative Analysis of Normalised Datasets *143*
5.3.7 PFL Viewer *148*
5.4 Technical Implementation 151
5.4.1 Business Intelligence Application *151*
5.4.2 Multidimensional Database *154*

5.5 Discussion 155
5.6 Distribution of Effort 158

CHAPTER 6: SPATIAL LOCALISATION & TURNOVER ANALYSES 159

6.1 Summary 159
6.2 Background 160
6.2.1 Experimental Design 162
6.3 Results 164
6.3.1 Protein Identification, Abundance and Subcellular Localisation 164
6.3.2 Determination of Protein Turnover 167
6.3.3 Distribution of Protein Turnover 169
6.3.4 Protein Turnover in Different Subcellular Compartments 171
6.3.5 Protein Characteristics Related to Turnover Rate 177
6.3.6 Protein Turnover and the N-terminal Amino Acid Rule 179
6.3.7 Amino Acid Frequency Distribution 180
6.3.8 Turnover Viewer 181
6.4 Technical Implementation 182
6.5 Discussion 183
6.6 Distribution of Effort 190

CHAPTER 7: PROTEIN ISOFORM, LOCALISATION AND TURNOVER ANALYSIS 193

7.1 Summary 193
7.2 Background 194
7.3 Quantification and Bioinformatics Analysis 198
7.4 Results 198
7.4.1 Protein Isoform Analysis: Candidate Approach 198
7.4.2 Protein Isoform Analysis: Rule of Thirds Approach 202
7.4.3 Protein Isoform Analysis: Three in a Row Approach 205
7.4.4 Isoform Analysis by Combined Protein Fractionation and Peptide MS 207
7.4.5 Correlating Post-Translational Modification with Protein Properties 213
7.5 Discussion 215
7.6 Distribution of Effort 219

CHAPTER 8: DISCUSSION 221

8.1 Proteomics Data Management and Analysis in the Future 227

APPENDICES 257

A. CAIRNGORM FRAMEWORK 257

B. METADATA DEFINITION 259

C. N-END RULE EVALUATION 263

D. RANDOM PROTEIN SAMPLING EVALUATION 269

E. LABTRACKER: IPAD BASED LABORATORY MANAGEMENT SOFTWARE 275

List of Figures

Figure 1: Protein primary, secondary, tertiary and quaternary structures. 35

Figure 2: Mass spectrometer overview. 37

Figure 3: Mass spectrometry instruments. 38

Figure 4: Typical SILAC experiment involving light and heavy labelled samples. 41

Figure 5: Mass spectrometry data workflow. 52

Figure 6: Abstract from Nature Biotechnology. 57

Figure 7: Egyptian astronomical measurement table. 61

Figure 8: Examples of charts available using the Google Chart API. 63

Figure 9: Screenshot of GapMinder (http://www.gapminder.org). 67

Figure 10: Relational versus dimensional database structure. 73

Figure 11: Snapshots of the NOPdb3.0 (http://www.lamondlab.com/NOPdb3.0/). 81

Figure 12: Entity-Relationship (ER) diagram for NOPdb3.0. 82

Figure 13: PepTracker workflow. 90

Figure 14: MsTrack workflow. 91

Figure 15: MS submission wizard interface. 92

Figure 16: Tag cloud generated from MS submission keywords. 95

Figure 17: PepTracker scheduler for handling data uploads. 98

Figure 18: DataVault browser views. 99

Figure 19: Early PepTracker interfaces created using Google Visualisation API. 101

Figure 20: DataVault graphs and charts. 102

Figure 21: Interactive DataVault interface. 103

Figure 22: PepTracker protein network map. 104

Figure 23: PepTracker customised protein network map. 105

Figure 24: Protein peptide alignment map. 106

Figure 25: Protein group information sheet. 107

Figure 26: PepTracker system architecture overview. 109

Figure 27: Overview of triple SILAC-based analysis of protein interaction partners. 128

Figure 28: Visualisation of contaminant profiles and threshold levels. 131

Figure 29: Protein Frequency Library construction and validation. 135

Figure 30: Filtering of PFL using experimental parameters (“dimensions”). 137

Figure 31: Application of PFL in the identification of specific protein interactors. 139

Figure 32: Normalisation of datasets using the PFL. 142

Figure 33: Analysis of protein interaction dynamics using normalised datasets. 144

Figure 34: The PFL Viewer tool. 150

Figure 35: PFL filter form within the PepTracker application. 151

Figure 36: Built-in normalisation functionality. 151

Figure 37: Sun diagram and logical model of SILAC data. 152

Figure 38: Pulse SILAC method. 163

Figure 39: Protein identification, abundance and subcellular localisation. 165

Figure 40: Distribution of protein turnover. 169

Figure 41: Protein turnover in subcellular compartments. 171

Figure 42: Distribution of protein turnover in subcellular compartments. 173

Figure 43: Subcellular clustering analysis of protein turnover. 175

Figure 44: Protein characteristics related to turnover rate. 177

Figure 45: PEST sequence analysis. 179

Figure 46: PepTracker spatial and turnover viewer. 181

Figure 47: Alternative splicing leading to protein isoforms. 196

Figure 48: NUDCD1 Protein isoform identification and localisation. 200

Figure 49: Protein isoform identification from protein sequence segmentation. 203

Figure 50: Protein isoform identification from consecutive peptide analysis. 206

Figure 51: Protein migration study on gel fractionation. 209

Figure 52: Protein intensity relation to gel slice 211

Figure 53: Protein isoform identification from gel fractionation 212

Figure 54: Phosphorylated proteins correlated with protein properties. 213

Figure 55: Phosphorylated post translation modification analysis with turnover. 214

Figure 56: Multidimensional analytics of protein properties. 231

List of Tables

Table 1: Successful Research Laboratory Software. 48

Table 2: Major Mass Spectrometry Data Analysis Software. 49

Table 3: Comparison of Major Mass Spectrometry Data Analysis Software. 50

Table 4: MaxQuant Output Files. 97

Table 5: Comparison of Peptide Data Quality for RNA Polymerase II Subunits. 145

Table 6: Embedding of Putative Specific Interaction Partners within Contaminants. . 147

Table 7: Average half-life for each amino acid at the first ten N-terminal positions... 264

Table 8: Average half-life for each amino acid at the last ten C-terminal positions.... 265

Table 9: Average turnover for each amino acid at the first ten N-terminal positions. 266

Table 10: Average turnover for each amino acid at the last ten C-terminal positions.267

Table 11: Analysis of amino acid occurrence of complete human proteome. 270

Table 12: Analysis of amino acid occurrence of turnover proteins. 271

Table 13: Analysis of amino acid occurrence of fastest turnover proteins. 272

Table 14: Analysis of amino acid occurrence of slowest turnover proteins. 273

Acknowledgements

Solely one person could not have completed the work within this thesis, rather it requires the support and collective encouragement of many.

First and foremost I would like to thank Professor Angus Lamond for believing in me and offering the opportunity to do this work. His supervision and continued support of my PhD has been unfaltering. His visionary approach to combining computing with life sciences has been inspirational and provided many ideas for this thesis work. His excellent leadership style has always been motivating and his constant encouragement and belief in my abilities has provided me with the platform to make the novel contribution described here. Without his backing, this truly would not have been the same PhD.

From the Lamond Laboratory, I would like to particularly mention Dr Severine Boulon and Professor Michel Boisvert for both providing me with the opportunity to work on exciting biological projects. Their patience and support cannot be underestimated and without them it would have been difficult to translate my computing skills to the field of life sciences research. They enthused me to become involved with the exciting biological happenings in the lab and were generous in providing me with the opportunity and supervision to work and gain experience at the lab bench.

I would like to especially mention Professor Peter Gregor, who acted as a co-supervisor and mentor throughout my PhD and provided much needed advice. Professor Mark Whitehorn and Andy Cobley from the School of Computing have also played a pivotal role in enabling me to combine proteomics with computing in new and exciting ways.

I am grateful to the Lamond Laboratory members, both past and present, for all their support, advice and friendship over the years. They are inspirational people who make Lamond Lab a unique place to learn and work. Thank you for teaching me proteomics and answering my many questions over the years. I have thoroughly enjoyed the opportunity to work in such an inspiring and motivating place.

I am grateful for the funding provided by the Biotechnology and Biological Sciences Research Council (BBSRC), which enabled me to do this PhD.

A big thanks to my family and friends for always believing in me and providing the support and understanding to make this PhD possible.

Candidate Declaration

I declare that I am the author of this thesis; that all references cited have been consulted by me; that the work of which this thesis is a record has been done by myself; and this thesis has not been previously accepted for a higher degree.

Yasmeen Ahmad

Executive Summary

Exponential data growth in life sciences demands cross discipline work that brings together computing and life sciences in a usable manner that can enhance knowledge and understanding in both fields. High throughput approaches, advances in instrumentation and overall complexity of mass spectrometry data have made it impossible for researchers to manually analyse data using existing market tools.

By applying a user-centred approach to effectively capture domain knowledge and experience of biologists, this thesis has bridged the gap between computation and biology through software, PepTracker (http://www.peptracker.com). This software provides a framework for the systematic detection and analysis of proteins that can be correlated with biological properties to expand the functional annotation of the genome.

The tools created in this study aim to place analysis capabilities back in the hands of biologists, who are expert in evaluating their data. Another major advantage of the PepTracker suite is the implementation of a data warehouse, which manages and collates highly annotated experimental data from numerous experiments carried out by many researchers. This repository captures the collective experience of a laboratory, which can be accessed via user-friendly interfaces.

Rather than viewing datasets as isolated components, this thesis explores the potential that can be gained from collating datasets in a “super-experiment” ideology, leading to formation of broad ranging questions and promoting biology driven lines of questioning. This has been uniquely implemented by integrating tools and techniques from the field of Business Intelligence with Life Sciences and successfully shown to aid in the analysis of proteomic interaction experiments.

Having conquered a means of documenting a static proteomics snapshot of cells, the proteomics field is progressing towards understanding the extremely complex nature of cell dynamics. PepTracker facilitates this by providing the means to gather and analyse many protein properties to generate new biological insight, as demonstrated by the identification of novel protein isoforms.

List of Publications

AHMAD, Y., BOISVERT, F. M., GREGOR, P., COBLEY, A. & LAMOND, A. I. 2009. NOPdb: Nucleolar Proteome Database-2008 update. *Nucleic Acids Research,* 37**,** D181-D184

BOULON, S.[*], **AHMAD, Y.**[*], TRINKLE-MULCAHY, L., VERHEGGEN, C., COBLEY, A., GREGOR, P., BERTRAND, E., WHITEHORN, M. & LAMOND, A. I. 2010. Establishment of a Protein Frequency Library and Its Application in the Reliable Identification of Specific Protein Interaction Partners. *Molecular & Cellular Proteomics,* 9**,** 861-879.

BOULON, S., PRADET-BALADE, B., VERHEGGEN, C., MOLLE, D., BOIREAU, S., GEORGIEVA, M., AZZAG, K., ROBERT, M. C., **AHMAD, Y.**, NEEL, H., LAMOND, A. I. & BERTRAND, E. 2010. HSP90 and its R2TP/Prefoldin-like cochaperone are involved in the cytoplasmic assembly of RNA polymerase II. *Molecular & Cellular Proteomics,* 39**,** 912-24.

TEN HAVE, S., BOULON, S., **AHMAD, Y.** & LAMOND, A. I. 2011. Mass spectrometry-based immuno-precipitation proteomics - the user's guide. *Proteomics,* 11**,** 1153-9.

BOISVERT, F. M., **AHMAD, Y.** & LAMOND, A. I. 2011. Chapter 20: The dynamic proteome of the nucleolus. *In:* OLSEN, M. O. J. (ed.) *The Nucleolus.* New York: Springer.

BOISVERT, F. M., **AHMAD, Y.**, GIERLINKSKI, M., CHARRIERE, F., LAMONT, D., SCOTT, M., BARTON, G. and LAMOND, A. I. 2011. A Quantitative Spatial Proteomics Analysis of Proteome Turnover in Human Cells. *Molecular & Cellular Proteomics,* doi:10.1074/mcp.M111.010728.

TEN HAVE, S., **AHMAD, Y.**, & LAMOND, A. I. 2011 Proteomics – Current Novelties and Future Directions. *J Anal Bioanal Techniques,* S3:001. doi:10.4172/2155-9872.S3-001.

---

[*] *co-authorship of paper.*

AHMAD, Y.[*], BOISVERT, F. M.[*] & LAMOND, A. I. 2011 Systematic analysis of protein pools, isoforms and modifications affecting turnover and subcellular localisation. *Molecular & Cellular Proteomics,* doi:10.1074/mcp.M111.013680.

Talks and Poster Presentations

AHMAD, Y. 19th February 2009 Bridging the Gap: Computing and Life Sciences. Invited Keynote Talk at GirlGeek Launch Scotland, Dundee, United Kingdom.

AHMAD, Y. 17th March 2009 Computational Tools for the Management & Mining of Cell Biology Data. Invited Talk for Master of Design and Design Ethnography Course, University of Dundee, Dundee, United Kingdom.

AHMAD, Y., & LAMOND, A. I. 5-7th May 2009 PepTracker. Poster at Plenary Meeting PROSPECTS, Majorca, Spain.

AHMAD, Y., & LAMOND, A. I. 26th September 2009 Computational Tools for the Management & Mining of Cell Biology Data. Poster at College of Life Sciences PhD Retreat, Kindrogan, United Kingdom.

AHMAD, Y., BOULON, S. & LAMOND, A. I. 3rd November 2009 Establishment of a Protein Frequency Library and its Application in the Reliable Identification of Specific Interaction Partners. Poster at Gene Regulation & Expression Symposium, University of Dundee, Dundee, United Kingdom.

AHMAD, Y. 02nd February 2010 Integrating Biological Research with Business Intelligence. Invited Talk for Master of Business Intelligence Course, University of Dundee, Dundee, United Kingdom.

AHMAD, Y. & LAMOND, A. I. 20th March 2010 Computational Tools for the Management & Mining of Cell Biology Data. Third Prize for Poster at College of Life Sciences Research Symposium, Crieff, United Kingdom.

AHMAD, Y., TEN HAVE., S. & LAMOND, A. I. 19-21st April 2010 Computational Tools for the Management & Mining of Cell Biology Data. Invited Talk and Poster at ProteoMMX: Strictly Quantitative, Chester, United Kingdom.

AHMAD, Y., BOISVERT, F., M. & LAMOND, A. I. 3rd November 2010 Protein Turnover & Spatial Viewer. Poster at 2nd Plenary Meeting PROSPECTS, Taormina, Sicily, Italy.

AHMAD, Y. 19th January 2011 Integrating Biological Research with Business Intelligence. Invited Talk for Master of Business Intelligence Course, University of Dundee, Dundee, United Kingdom.

AHMAD, Y., BOISVERT, F., M. & LAMOND, A. I. 26th March 2011 A Quantitative Spatial Proteomics Analysis of Proteome Turnover in Human Cells. Third Prize for Poster at College of Life Sciences Research Symposium, Crieff, United Kingdom.

AHMAD, Y. & WHITEHORN, M. 15th April 2011 University Uses Business Intelligence Software to Boost Gene Research. Invited Keynote Speech at Microsoft SQL Server 2008 R2 Launch, London, United Kingdom.

AHMAD, Y. & WHITEHORN, M. 16th April 2011 Denormalisation. Invited Talk at SQLBitsVI, Westminster, London, United Kingdom.

AHMAD, Y., BOISVERT, F., M. & LAMOND, A. I. 24th May 2011 A Quantitative Spatial Proteomics Analysis of Proteome Turnover in Human Cells. Poster at MaxQuant Summer School, Max-Planck Institute for Biochemistry, Munich, Germany.

AHMAD, Y., BOISVERT, F., M. & LAMOND, A. I. 10th June 2011 Quantitative Peptide Level Analysis of Protein Isoforms in Human Cells. Poster at 2nd PICLS Annual Symposium, University of Dundee, Dundee, United Kingdom.

AHMAD, Y., BOISVERT, F., M. & LAMOND, A. I. 14th June 2011 Unlocking the Potential of Proteomics Data. Talk at Divisional Seminar Series, University of Dundee, Dundee, United Kingdom.

AHMAD, Y. & LAMOND, A. I. 11th July 2011 Super Experiments: The Future of Proteomics and Cell Biology. Poster at Wellcome Trust Centre for Gene Regulation & Expression: Governer's Visit, University of Dundee, Dundee, United Kingdom.

AHMAD, Y. 25th August 2011 Introducing LabTracker – molecular biology meets the iPad. Talk at Mini-PROSPECTS Meeting, Karolinska Institute, Stockholm, Sweden.

AHMAD, Y. & LAMOND, A. I. 4th October 2011 Systematic analysis of protein pools, isoforms and modifications affecting turnover and subcellular localisation. Poster at Gene Regulation & Expression Symposium, University of Dundee, Dundee, United Kingdom.

Abbreviations

1NF/2NF/3NF	First/Second/Third Normal Form
AJAX	Asynchronous JavaScript & XML
API	Application Programming Interface
BI	Business Intelligence
CRUD	CReate, Update and Delete
CSV	Comma Separated Variable
DMEM	Dulbeccos's Modified Eagle Medium
DNA	DeoxyriboNucleic Acid
DPMDB	Global Proteome Machine DataBase
EBI	European Bioinformatics Institute
ER	Entity Relationship
ETL	Extract, Transform and Load
GFP	Green Fluorescent Protein
GUI	Graphical User Interface
GO	Gene Ontology
GWT	Google Web Toolkit
HCI	Human Computer Interaction
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
HUPO	HUman Proteome Organisation
IDE	Integrated Development Environment
IPI	International Protein Index

KD	Knowledge Discovery
LDAP	Lightweight Directory Access Protocol
LIMS	Laboratory Information Management System
MDX	MultiDimensional eXpressions
MIAPE	Minimal Information About a Proteomics Experiment
MS	Mass Spectrometry
MVC	Model View Controller
NOPdb	Nucleolar Proteome DataBase
OLAP	OnLine Analytical Processing
PFL	Protein Frequency Library
PICR	Protein Identifier Cross-Referencing
PRIDE	PRotein IDEntifications
PSI	Proteomics Standards Initiative
PTM	Post Translational Modification
RDBMS	Relational DataBase Management System
REST	REpresentational State Transfer
RIA	Rich Interactive Application
RNA	RiboNucleic Acid
SDK	Software Development Kit
SILAC	Stable Isotope Labelling by Amino acids in Cell culture
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
SVG	Scalable Vector Graphics

TPP	Trans-Proteomic Pipeline
URL	Uniform Resource Locator
WebGL	Web Graphics Library
WfMC	Workflow Management Coalition
WMS	Workflow Management System
WYSIWYG	What-You-See-Is-What-You-Get

Introduction

Biology is a broad ranging field, spanning the study of whole living organisms to the molecular scale of the cell, hence defining a variety of sub disciplines. Cellular biology focuses on examining the building blocks of life in order to further understand the structure, function, growth, origin, evolution, distribution and taxonomy of life. Thanks to technical and experimental innovations it is now possible to make high throughput, quantitative measurements of cellular DNA, RNA and protein molecules on an unparalleled scale. This allows researchers to design experiments in new ways that promise major insights into the mechanisms of cell growth and the relationships between gene function and human disease.

The main role of genes is to instruct a cell on how to make the different types of proteins that control cellular metabolism and make up the fabric of subcellular structures. Whereas biologists previously concentrated on studying one or two proteins in isolation, it is now possible to detect and measure changes in the levels and properties of thousands, or even tens of thousands, of genes and their protein products in a single experiment. This field is known as “proteomics” or “functional genomics”.

Proteomics is continuously evolving, with better instrumentation available year on year and improvements in experimental techniques and protocols. Laboratories now have easier access to mass spectrometry instrumentation and are able to carry out high throughput research. The corollary of large-scale modern proteomics is resulting in the generation of big volumes of data that represent the quantitative measurements of proteins residing in cells. For example, current experiments can already generate in the order of 600GBytes of raw data and it is envisioned future experiments will generate even larger datasets at a more granular level.

This exponential increase in the volumes of mass spectrometry data is not limited to the proteomics domain. Data growth has been experienced across many fields, including genomics, complex physics simulations, finance and retail. This trend of larger datasets has raised issues surrounding the capture, management and processing of data within acceptable time frames and is now termed as the field of ‘Big Data’. Big data sizes are a constantly moving goalpost, with current big data being ranging from a

few dozen Terabytes to many Petabytes. This definition puts proteomics in the arena of Big Data and opens up the possibility of using Big Data technologies to help manage data sets in a collective manner.

Existing tools in the proteomics domain, such as the Trans Proteomic Pipeline, have made significant efforts to make the analysis process of mass spectrometry data more straightforward. However, to date, none of the software' created provide data warehouse and complex downstream analytic capabilities for quantitative proteomics data. Furthermore, many software focus on a single user, single machine workflow, which is restrictive in terms of the size of datasets that can be handled and fails to take advantage of the experience captured in previous datasets.

Traditionally, biologists perform experiments serially and analyse the resulting data from a single experiment largely in isolation from other datasets. Primarily researchers will concentrate on the follow up analysis of only a small subset of the resulting data, resulting in most of the obtained information being effectively discarded and its value lost. In the absence of dedicated software tools, designed specifically to handle proteomics data, the value of information generated is further reduced due to limitations with the available analysis tools (e.g. Excel), which are not suited to the scale and complexity of these new types of data.

Furthermore, the data are often not annotated with well-structured and comprehensive metadata describing experimental protocols applied during generation, which limits the amount of automated analysis that can take place across datasets. If datasets could be collated and stored, it is envisioned that all of the data generated in every experiment could be collectively used to aid in future analysis, rather than discarding them as irrelevant based on the narrow interests of the researcher seeking to test a specific, often narrow hypothesis.

Another challenge resulting from the lack of suitable software tools has been that the experimentalists, who best understand the design and meaning of their studies, however can not perform the analysis of their data because they lack the required computing skills. Instead, the large datasets are passed over to specialised Biocomputing groups, who do not perform experiments themselves, for analysis and interpretation. This separation of data analysis from data generation can lead to

misunderstandings on both sides and can complicate and hinder an efficient knowledge discovery process. It can also contribute to a rather piecemeal approach for analysing biological data. Software created by specialist Biocomputing groups can be difficult to setup and utilise, as these software are often not designed for non-computational experts. This can lead to frustration on the part of researchers whose expertise lie in biology rather than computing.

In order to bridge this gap between life sciences and computing, well-designed interfaces are required that can hide the underlying complexity in algorithms and computation, to provide researchers with the ability to carry out their own analysis. In order to implement such interfaces a user-centred design philosophy can be employed which focuses on the needs, wants and limitations of end users. At each stage of the software development, specific focus is placed on involving users in design and usability evaluations. The feedback from informal interviews, focus groups and one-to-one discussions can then be fed back into the software development cycle to improve the outcome. This increases the likelihood of user acceptance and aids in the identification of problems and issues early, which would otherwise take a lot of time and effort to resolve later in the design and development process.

By focusing on users, the opportunities arising from the challenges in proteomics have been identified and embraced to build a new approach for extracting maximum value from proteomics data. It has been recognised that a custom data environment must be developed for management and analysis of quantitative proteomics data. The aim of this is twofold; first to ensure that researchers can make best use of all data resulting from every experiment carried out in the laboratory and second, to place the ability to carry out sophisticated data analysis and mining back into the hands of the experimentalists who generate the data.

This thesis describes the development of new software tools that facilitate the ability of researchers to manage and intuitively analyse their own data and convert it into information. This in turn translates to biological knowledge and understanding, which not only furthers the field but can also be reapplied to follow-up experiments and used to design and test hypotheses.

Chapter 1: Literature Review

Chapter 1 provides a background to cell biology and proteomics, describing the use of mass spectrometry and various experimental techniques as well as describing the challenges faced by the field in moving forward (section 1.1). These challenges can be summarised as data management, visualisation and analysis (sections 1.2-1.4).

1.1 Cell Biology & Proteomics

1.1.1 Genomics & Proteomics

The Human Genome Project was a major international endeavour aimed at identifying and mapping all genes, which control hereditary characteristics in living organisms. A gene is any given segment along a DNA strand that encodes instructions allowing a cell to produce a specific product - typically a protein. However, since the completion of the human genome project (2003) (Lander et al., 2001, Venter et al., 2001) the focus of research has changed from working at the genome level, identifying and mapping genes, to documenting the function of genes and realising how changes in the sequence relate to health and disease at the cellular level. Genes, while holding the code to build proteins, are functionally non-descript, as the regulation and expression of the gene is a result of proteins in the cell reacting to environmental and chemical stimuli. Additionally, splicing and processing and modification variants, which arise at the protein level, may alter the structure and function of the proteins from a given gene. By researching the proteins expressed by genes under various conditions, the field of life sciences has made significant contributions to the understanding of how the human body functions. This research has led to the definition of a new field: proteomics, which aims to discover, annotate and describe the properties of proteins in living organisms. The field of proteomics is vast, covering the analysis of all proteins encoded by a genome. Typically there are multiple proteins associated with each gene, resulting in the total number of protein products extending far beyond the estimated 20,000-25,000 genes in the human genome.

The molecular basis of genes is DNA (DeoxyriboNucleic Acid), which consists of long polymer chains that are composed of four subunits, known as nucleotides or bases. These four nucleotides are: adenine (A), cytosine (C), guanine (G), and thymine (T). The order in which the nucleotides appear in a DNA strand determines the biological information encoded by the strand. The chemical attractions between the nucleotides

result in the appearance of molecule pairs, such that adenine always pairs with thymine and guanine pairs with cytosine. This base-pairing characteristic makes it straightforward to work out the complementary DNA strand of any single-stranded DNA sequence. These two DNA strands together form a structure with the well-known double-helix shape and tightly bind to form chromosomes. It is this DNA structure that forms the physical basis of inheritance in a living organism.

When a cell reproduces it divides into multiple cells, passing on genetic information via DNA strands. It is these DNA strands that are used by the cell during protein synthesis, the process by which cells build new proteins. This process requires the use of DNA from within the cell nucleus. Each DNA strand within a chromosome contains genes, which each hold a genetic ‘blueprint’ required to build a protein molecule. Depending on which genes are active or inactive will directly influence the proteins that are synthesised within a cell.

Similar to DNA and RNA, proteins are linear structures, which can be represented by a text string. Whereas DNA and RNA use a four-character alphabet, proteins require a twenty-character alphabet to represent each possible amino acid in a protein sequence. When amino acids join together, they form polypeptides and each protein can consist of one or more polypeptides. The amino acid sequence of a protein can vary in length from typically 50 to 2000 amino acid residues and is known as the primary structure. The attraction between amino acids can cause a protein primary sequence to coil or fold, forming a secondary structure. Finally the tertiary structure describes the overall 3D protein molecule, which is the secondary structure further folded back on itself (see Figure 1).

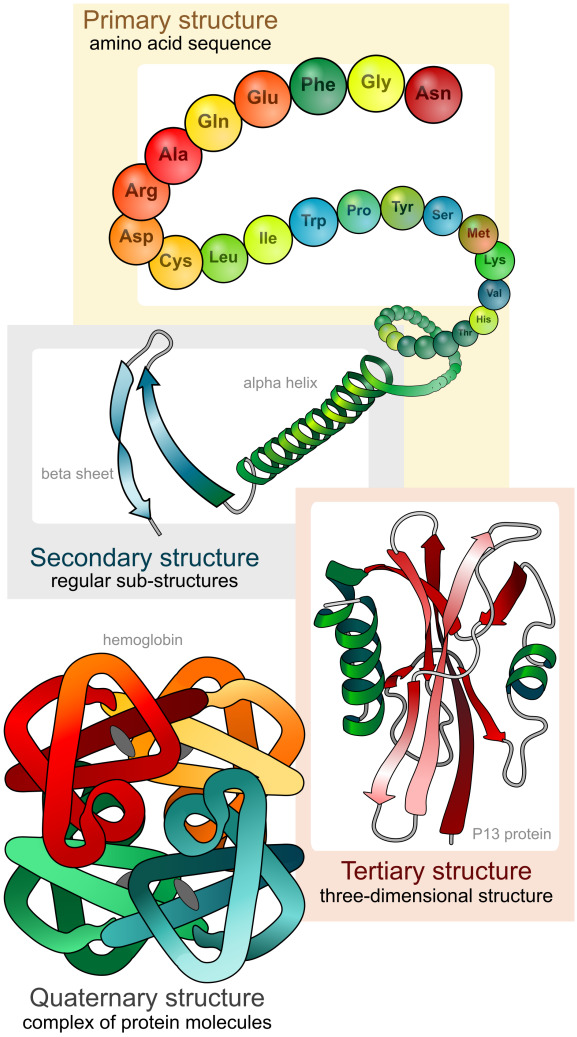


Figure 1: Protein primary, secondary, tertiary and quaternary structures.

(http://en.wikipedia.org/wiki/File:Main_protein_structure_levels_en.svg)

During protein synthesis, or directly following synthesis, proteins can undergo a process of posttranslational modification. This modification is a chemical alteration of the protein, which can alter the physical and chemical properties, folding, stability, activity and function of the protein.

Whilst genetic information is static, the protein complement of a cell is dynamic. Proteins play a pivotal role in the life of an organism, they are the elements within a cell that provide structure, produce energy, as well as allow communication, movement, and reproduction (Cho, 2007). Proteins are the main macromolecules of an organism hence, when studying an organism biologists usually either investigate proteins, or molecules that have been made from, or by, proteins. Due to the pivotal role played by proteins in providing the basic functional and structural framework for

cellular life, it is imperative for researchers to be able to understand the life cycle of proteins.

The study of functional and behavioural proteomics aims to answer questions such as which form of a protein has been found in a cell (isoform) (Neubauer et al., 1998, Rappsilber and Mann, 2002b, Ahmad et al., 2011), how much of a protein is in a cell (intensity and abundance), where it is located in a cell (spatial localisation), who it communicates with (interactions), what structural conformation it may have (structure), binding surfaces or locations (Nett et al., 2010, Rappsilber, 2011) and what chemical modifications it may gain or lose. Protein characterisation has allowed investigation of potential pathogen species (Wilson et al., 1999). These questions can be answered through a variety of experimental techniques. Localisation of proteins can be studied through cellular fractionation techniques that allow measurement of proteins within different sub-cellular compartments (Boisvert et al., 2010, Boisvert et al., 2011). Protein-protein interactions can be studied using methods such as immunoprecipitation (Trinkle-Mulcahy et al., 2008c, Boulon et al., 2010a, Boulon et al., 2010b), involving the use of an antibody bound to the protein of interest to isolate interaction partners, or cross linking (Rappsilber et al., 2000, Maiolica et al., 2007, Rappsilber, 2011), which covalently binds together interacting proteins. It has been shown that these approaches to studying protein-protein interactions also require adequate bioinformatics software (Maiolica et al., 2007). Furthermore, protein modification can be identified via enrichment of modifications within a biological sample and novel in-machine selection processes to analyse specific ions which characterise peptide modifications (Macrae and Ferguson, 2005, Vertegaal et al., 2006, Matic et al., 2008, Westman et al., 2010, Ahmad et al., 2011).

The Lamond Laboratory, based in the Wellcome Trust Centre for Gene Regulation & Expression, is playing a prominent role in the development and application of new quantitative and high throughput methods for the analysis of gene expression and cell biology. In particular the Lamond group focus on how cancer and other diseases can result in changes in the spatial distribution, stability and function of proteins in human cell lines. Rather than analysing the cellular proteome as a static snapshot in time, advanced studies are now looking at the function and dynamics of proteins on a proteome-wide scale involving large-scale use of proteomics technologies. The most

accurate and high throughput method for protein analysis used in science today is Mass Spectrometry (MS).

1.1.2 Mass Spectrometry

Mass Spectrometry relies on highly accurate mass measurements of peptide preparations - typically tryptic digests of proteins, where a protein sequence is cut after every Lysine and Arginine amino acid resulting in computationally predictable pieces of protein, namely peptides. These pieces of protein are measured in a mass spectrometer, which distinguishes between ions based on mass-to-charge (m/z) measurements. A mass spectrometer has three main components: an ion source, m/z analyser and a detector (see Figure 2).

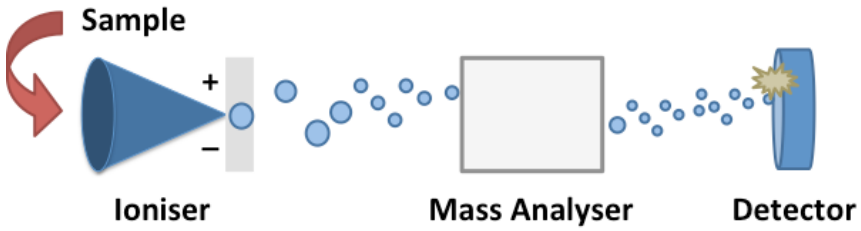


Figure 2: Mass spectrometer overview.

In order to make mass-to-charge measurements of ions, each molecule is ionised and the ion is propelled into a mass analyser by an electric field. In order to allow accurate measurements to be taken, the ions are released from the ioniser over a period of time. Popular types of ionisation include electrospray ionisation (ESI) (Fenn et al., 1989) and matrix-assisted laser desorption/ionisation (MALDI) (Karas and Hillenkamp, 1988), due to the little or no fragmentation of the molecules that occurs during the ionisation and desorption process. When the ions reach the mass analyser they must be sorted depending on their mass-to-charge ratio. There are two main types of mass analyser: those that use an electric field and those that employ the use of magnetic fields. There are four commonly used mass analysers: ion trap, time-of-flight (TOF), quadrupole and Fourier transform ion cyclotron (FT-MS) analysers (Aebersold and Mann, 2003). The choice of analyser depends on a number of factors related to the design and performance of each. The analysers can be used individually or placed in tandem with one another to take advantage of the strengths of each. Finally, the detector (usually a specific type of electron multiplier) takes readings by producing an

electric charge each time it is struck by an ion. The goal of the detector is to record the abundance of each mass-to-charge ion.

The Fingerprint Proteomics Facility and the Lamond Laboratory, in the College of Life Sciences, have a variety of instruments from Thermo Scientific Fisher. These include the LTQ Orbitrap XL, LTQ Orbitrap VELOS and Q Exactive (see Figure 3). The Orbitrap instruments are based on technology developed by Alexander Makarov. They include a special type of ion trap where ions are electrostatically trapped in an orbit around a central, spindle shaped electrode. The electrode confines the ions so that they both orbit around the central electrode and oscillate back and forth along the central electrode's long axis. This oscillation generates a current in the detector plates, which is recorded by the instrument. The frequency of the oscillations is directly dependant on the mass-to-charge of the ions.

Figure 3: Mass spectrometry instruments.

The output of the mass spectrometer is a spectrum that can be used to match detected ions to masses of predicted tryptic peptide sequences, which can be calculated due to the known chemistry of amino acids. High throughput technologies, such as MS, have made it possible for researchers to identify hundreds of thousands of peptides, leading to thousands of protein identifications in a single analysis. This is known as ‘bottom up’ proteomics as the process works from the peptide level back up to the complete protein.

The great advance seen in the mass spectrometers of today comes from many innovators. Even though MS is now popular in cell biology its roots lie in physics, where it was first conceived by inventor Joseph John Thomson who received a Nobel Prize for Physics in 1906 for his work on the existence and properties of ions (Thomson, 2010). Francis Aston carried on the work of Thomson and won his own Nobel Prize in 1922 for

Chemistry and is credited for building the first mass spectrometer that could measure the masses of atoms (Squires, 1998). The instrument was then modified and continuously improved until the end of the 1930's by which time mass spectrometry had become an established technique for the separation of atomic ions by mass. However, there still remained the challenge of moving large molecules into the gas phase without extensive fragmentation and decomposition. The next major development came in the form of Wolfgang Paul's invention of the quadrupole and quadrupole ion trap (Paul, 1990), which earned him the Nobel Prize in 1989 for Physics. Then, in 1988, the ESI and MALDI ionisation techniques appeared almost simultaneously and revolutionised biological MS. John Fenn received a shared 2002 Nobel Prize in Chemistry for his development of ESI (Fenn, 2002) along with Koichi Tanaka for his development of a laser desorption method of protein ionization (Tanaka et al., 1988). Both of these techniques opened up the world of mass spectrometry analysis to biological macromolecules, such as proteins.

Mass spectrometry-based proteomics can now be applied to samples from various model organisms, including human (Nilsson et al., 2010), yeast (de Godoy et al., 2008), nematodes (Larance et al., 2011), trypanosomes (Nett et al., 2009b, Nett et al., 2009a) and drosophila (Brunner et al., 2007). Multiple samples from both the same organism and different organisms can be compared using mass spectrometry combined with stable isotope labelling. This involves labelling samples with different stable isotopes, after which the proteins and peptides in different samples are still chemically (almost) identical but have a different mass. This shift in mass can be used to compare the same proteins in multiple samples of generated under different conditions.

There are many isotope-labelling methods, the most popular techniques include:

- Isotope Coded Affinity Tags (ICAT) (Gygi et al., 1999a) is one of the first methods developed to differentially label peptides. This technique uses a reagent that can carry either a light or heavy tag that covalently attaches to a protein. Labelling two samples with either the light or heavy tag allows comparison between two conditions. The samples are then combined, before digesting and analysing by mass spectrometry. ICAT has a dependence on the occurrence of cysteine's in the proteins of interest and the number of cysteine

residues as this can negatively affect the sequence coverage of a protein identification.

- Isobaric Tags for Relative and Absolute Quantification (iTRAQ) (Ross et al., 2004) uses covalent labelling of N-terminus side chains of peptides with tags of varying size. This approach differs from ICAT as peptides are fractionated before labelling takes place. Final database searching on the output of the mass spectrometry can match fragmentation data of the peptides and tags to identify and quantify the peptides and hence proteins from which they originate. An advantage of iTRAQ over ICAT is that protein sequence coverage is higher as all peptides are labelled and available for analysis.
- Stable Isotope Labelling By Amino Acids (SILAC) (Ong et al., 2002) differs from both iTRAQ and ICAT as it metabolically labels proteins via incorporation of labelled arginine, lysine or both. This gives rise to the possibility of four different samples, also considering an unlabelled sample. This tag less approach provides benefits over other labelling strategies as it alleviates potential issues with liquid chromatography, for example co-elution, and reduces proteomic complexity, as the cells analysed are usually homogenous. This approach is described in detail in 1.1.3 Stable Isotope Labelling using Amino Acids in Cell Culture.

1.1.3 Stable Isotope Labelling using Amino Acids in Cell Culture

Advanced techniques, including MS combined with Stable Isotope Labelling by Amino acids in Cell culture (SILAC) (Ong et al., 2002), are being used to not only accurately identify but also quantify proteins within biological samples. SILAC protocols obtain quantitative measurements through comparison of light and heavier forms of the same peptide, arising from the presence of heavier, stable isotopes such as ^{13}C, ^{2}H and ^{15}N. These stable isotopes are incorporated into proteins by in vivo labelling, i.e. growing the cells in specialised media where specific amino acids, typically arginine and lysine, are replaced with corresponding heavy isotope-substituted forms in which either all carbons, or combinations of carbon's, hydrogen's or nitrogen's are isotope-labelled. This labelling of peptides, using isotopes, produces a mass shift, which can be measured by the mass spectrometer and appears in the mass spectra. SILAC not only aids with quantification of proteins but also allows researchers to compare samples using differential labelling.

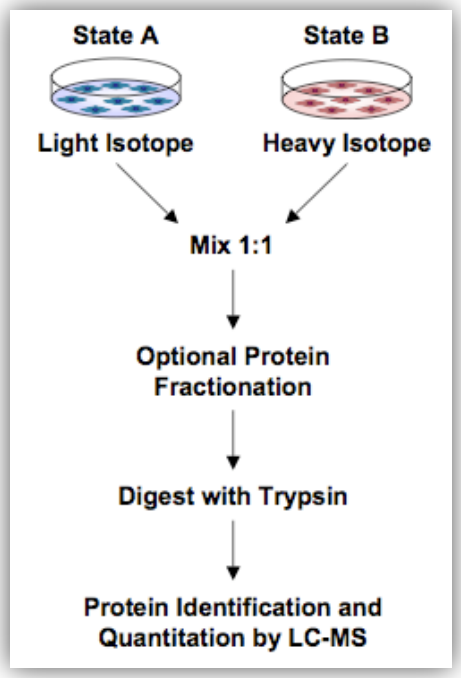


Figure 4: Typical SILAC experiment involving light and heavy labelled samples.

The advent of techniques such as SILAC, combined with the advancements in instrumentation have allowed the use of mass spectrometry not only for protein identification and study of modifications, but also made it possible to study quantitative expression of proteomes (Rappsilber and Mann, 2002a, Ong and Mann, 2005).

1.1.4 Challenges in the Proteomics Field

When attempting to identify proteins from peptide identifications, there are a number of problems faced by researchers. Theses challenges are two fold, related to either the experimental method or the computation. The experimental problems can be summarised as:

- difficulty in detection of low-abundance proteins due to the large dynamic range of protein abundances found in cells,
- loss of proteins during enrichment strategies due to their fragility,
- the volatile nature of some proteins in certain environments, making them highly unstable,
- selective MS/MS, which is pre-set to usually the top 10 highest intensity ions, resulting in not all peptides being selected for MS/MS and placing bias towards high abundance peptides, and

- the lack of reproducibility between separate mass spectrometry runs.

Due to these problems, and the range of tasks and objectives in proteomics, it has been stated that finding a universal technology "that enables you to identify all proteins, in all samples, all of the time" is near impossible (Kevin Auton, Chief Executive, Proteomics Technology Developers, NextGen Sciences of Huntingdon, UK) (cited in (Gershon, 2003)). Due to this, over recent years, much emphasis has been placed on trying to make scientific protocols and experimentation more accurate through advancements in instrumental, experimental and computational methods.

Using improved scientific protocols coupled with MS, biologists are now able to generate larger volumes of data related to a greater number of proteins and peptides. The improvements in MS equipment make it possible to sample more ions, resulting in a greater number of spectra obtained per unit of time. Currently each mass spectrometer is able to generate hundreds to tens of thousands, of fragment ion spectra per hour of data (Nesvizhskii et al., 2007). However, scientists are not yet able to benefit fully from the discoveries that can be made from these data, due to the computational and statistical challenges that have arisen. These problems can be summarised as:

- cost of data storage,
- data management issues that are simply not addressed by storing data but require storage of metadata related to the raw data,
- MS/MS spectral algorithms mostly report a significance of a match, these algorithms can not report an absolute measure, which means there is ambiguity in the interpretation of the output and confidence can only be gained through finding the same protein identification in multiple, separate MS experiments,
- large numbers of spectra are difficult to combine and analyse, resulting in experiments being analysed individually rather than part of a larger goal,
- complexity of the proteomics experiments resulting even more complex datasets that are impossible to visualise and analyse with existing, popular scientific tools such as Excel,

- the lack of consensus on data formats and data standards leading to problems with sharing of data,
- majority of MS/MS spectra represent the noise or minor contaminants and therefore have no value to the aim of the experiment and should not be included within the analysis process,
- contamination and noise detected between samples is likely to overlap, however to build a baseline a large number of experimental datasets would have to be compared for which there is no current infrastructure and
- existence of proteins as multiple isoforms or post translational states make it difficult to experimentally and algorithmically detect and differentiate between these.

In the early days, the main goal of MS based proteomics may have been to identify and analyse proteins faster, with increased sensitivity and reliability, however this focus has changed. On its own MS is inefficient, especially in the processing of data (Aebersold and Goodlett, 2001). An appropriate proteome analysis platform is required to convert data, obtained during experiments, into useful knowledge that can be applied to make biological discoveries of significance.

It has already been shown that combining computing with biology in the field of bioinformatics can be very useful. The creation of niche tools combined with databases can help in the analysis of mass spectrometry data within a variety of situations, including the investigation of structural proteomics (Rappsilber, 2011) and post-translational modifications of proteins (Martin et al., 2010).

1.2 Data Management Technology

1.2.2 Current Data Management in Life Sciences

It has been recognised that manual methods of inspection and analysis are no longer viable and instead tools are required for computational analysis of these MS data (Kumar and Mann, 2009). To date there is no method to routinely capture, manage and archive datasets from such studies. A proteomics consultant, Sara Ten Have, has described this problem: “despite our reliance on computation, most scientists are not capable of complex data storage and analysis computing, and therefore rely on computer programmers to do this for us”. Advances in technology have allowed for

more sophisticated proteomics experiments, which have resulted in the generation of an increased volume and complexity of data that demands the development of new tools due to the non-existence or inadequacy of current tools, such as Excel. In these situations, biological researchers have been forced to carry out minimal analysis manually and then hand-over their datasets to bioinformaticians (typically external to the laboratory) who have the necessary computing skills to handle these data. This is frustrating for the biologists who are experts in how the data are generated and who understand the meaning and limitations of the data. Having driven the formulation of the initial hypothesis that led to the experiments and data generation, the biologists are more acutely aware of how they would like to question the data further and its potential. Furthermore, there can be limited contact between bioinformaticians and biologists resulting in minimal information exchange regarding the context of the data and the processes involved in generating the data, potentially leading to errors in analysis.

Despite the many development projects that have taken place by bioinformaticians and software developers, an evident gap remains in the uptake of new software solutions. Furthermore, some systems that were developed for MS-based proteomics are no longer maintained or available, including:

- Rosetta Biosoftware (support discontinued in July 2011) – this technology and corresponding assets were bought by Microsoft in 2009 and sale of the software was discontinued. Support for existing customers was stopped in July 2011. Microsoft used the assets to drive progress on their own Microsoft Amalga Life Sciences platform.
- PeptideSearch (last update was September 2007) – this peptide search software was created by the Mann group while he was at the Bioanalytical Research Group in EMBL-Heidelberg. Once Matthias Mann left this research institute the software was left to be maintained by the next group leader but is no longer maintained or available.
- Sherpa (last reported release in 2000) – this software was developed by the Ken Walsh Laboratory (Biochemistry Department, University of Washington, Seattle, WA) to aid in the correlation and interpretation of LC/MS and MS/MS

spectra, however it was never developed beyond a beta release. The programmer discontinued development and support of the software.

The main factors attributing to software becoming unavailable include:

- software developers moving onto new projects or moving institutes,
- unusable nature of many systems,
- software becoming out-dated,
- inadequate support for end users,
- complex nature of the science resulting in software that does not meet requirements,
- software groups being bought over by companies,
- new instrumentation,
- evolving standards (see 1.2.9 Data Standards) and
- high licensing costs of commercial software.

Researchers, with a background in life sciences, often come across complex analytical needs, which drive them to create custom software in laboratories. This results in the researcher choosing to learn the necessary computing skills that can help them build scripts, which later become tools. However, this strategy often results in these tools becoming out of date and unavailable due to the researcher moving onto another institute and leaving behind software that is no longer maintained.

On the other hand, successful software projects, like the Rosetta Biosoftware, have suffered due to the success of their software developers. The Rosetta Biosoftware was bought over by a company (Microsoft), who wanted to access the talent of the software developer team for other projects, which led to the software being discontinued.

The success of a piece of software is not always attributed to its functionality, but is also dependant on how this functionality is implemented. To use many of the tools that are currently available on the market, computer expertise is required (Mead et al., 2007). Scientists trained in biology generally do not hold such expertise and should not be expected to learn complex computer skills to be able to use the tools. Instead, the onus should be on developers to create usable, accessible software that can be utilised

with minimal effort from researchers. Often developers neglect the interface and hence neglect to understand the importance of effective design in the success of software.

Not only should software be usable, it should also be supported by adequate user documentation, video tutorials and/or workshops. As the functionality of a piece of software increases, it inherently becomes more complex and so user documentation and support becomes imperative. Many successful projects, such as MaxQuant and Skyline, have created user groups. These user groups aid in establishing a community of users who can help one another and answer questions, removing the pressure on developers to answer all questions. Furthermore, having workshops at conferences and holding dedicated user group gatherings, like the MaxQuant Summer School, provide opportunities for users to speak with developers and gain hands-on training.

Integrating computing with 'wet' experiments is another challenge that is often underestimated. The complex nature of the research, carried out by biologists, is difficult to communicate and can act to hinder collaborations between the two disciplines. Establishing effective communication is pertinent, as a software development project cannot succeed without effective two-way communication that allows focused and well articulated requirements to be elicited and implemented.

Another major challenge for laboratories is the high licensing costs of software, especially for tailor made applications. As seen from Table 3, software entails multiple costs in terms of purchase and paying for support, which often comes as an additional, yearly subscription fee. Even if a laboratory decides to opt for an open-source software package, they are then reliant on the software developers to maintain the solution and keep it up-to-date and compatible with other analysis software that the researchers either must, or may want to use in parallel.

Where there is an evident gap in the market for software with novel functionality, there is a case for biological research laboratories to commission research and to create tailor made software tools that can meet the demands of their research in its advanced state. However, it should be understood that under these circumstances, considerable thought and effort is required to establish a team of software developers that will be responsible for implementing functionality, as well as continuing to

maintain and support the software. Table 1 lists successful software developed in research laboratories. From the details of the software development, it is evidenced that these laboratories view this software as a major resource that requires significant development effort. Furthermore, these software are supported through multiple methods, including videos, tutorials and workshops.

Software	Details	
DTASelect Contrast	*Laboratory:*	Yates Laboratory (http://fields.scripps.edu/researchtools.php)
	Aim:	DTASelect organizes and filters SEQUEST identifications, reducing the time required to interpret the results for each sample. Contrast differentiates multiple samples and comprises a powerful meta-analytical tool.
	Developers:	2
	Support:	Tutorial, Manual, Email Address
	Cost:	Freeware
MaxQuant	*Laboratory:*	Mann Laboratory (http://maxquant.org/)
	Aim:	*Quantitative proteomics software package designed for analysing large mass-spectrometric data sets.*
	Developers:	6
	Support:	Written Tutorial, Workshops, Forum
	Cost:	Freeware
PeptideProphet ProteinProphet	*Laboratory:*	Aebersold Laboratory, Institute for Systems Biology (http://peptideprophet.sourceforge.net/) (http://proteinprophet.sourceforge.net/)
	Aim:	Automatic validation of peptide assignments from database search programs, such as SEQUEST, to MS/MS spectra.
	Developers:	4
	Support:	Video Tutorial, Written Tutorial, Workshops, Forum, Mailing List
	Cost:	Open Source
SkyLine	*Laboratory:*	MacCoss Lab Software (https://skyline.gs.washington.edu/labkey)
	Aim:	*Windows client application for building Selected Reaction Monitoring (SRM) / Multiple Reaction Monitoring (MRM) and Full-Scan (MS1 and MS/MS) quantitative methods and analysing the resulting mass spectrometer data.*
	Developers:	9
	Support:	Video Tutorials, Written Tutorials, Workshops, Forum
	Cost:	Freeware
Various Software Tools, e.g. DanteR, DeconTools, DeconMSn, MultiAlign, VIPER	*Laboratory:*	Pacific Northwest National Laboratory (http://omics.pnl.gov/software/)
	Aim:	Multiple tools to aid with the analysis of mass spectrometry data, including Fasta File/Protein Sequence/Protein Database related tools, MS & MS/MS Analysis, Data Analysis & Data Presentation, MS Data File Utilities and Mass Spectrometry Auxiliary tools.
	Developers:	40+
	Support:	Written Tutorials/Help Guides, Workshops
	Cost:	Open Source

Table 1: Successful Research Laboratory Software.

In research laboratories, not all requirements will be well defined and, furthermore, these requirements will continuously evolve. Hence the task of creating software should be viewed as an on-going effort as opposed to a short-term project.

For successful uptake, software development in proteomics requires the implementation of novel software that is automated, robust and user-friendly (Gershon, 2003). The above challenges should all be considered to ensure a sustainable solution.

1.2.3 Comparison of Existing Software

Shown below is a table summarising the main downstream analysis software available for analysing quantitative mass spectrometry datasets. These software contain a suite of tools that work collectively to allow full analysis of mass spectrometry datasets. There are many other tools developed by research laboratories that have not been included in the comparison below as these tools carry out specific tasks that are not comparative to enterprise level applications. Table 3 shows a review of the different software.

Software	Supplier
MaxQuant	Mann Laboratory http://maxquant.org/
PepTracker	Lamond Laboratory http://www.peptracker.com/moreInformation/
Progenesis	Nonlinear Dynamics http://www.nonlinear.com/products/progenesis/lc-ms/overview/
ProteinScape	Protagen http://www.proteinscape.com/
ProteoIQ	BioInquire http://www.bioinquire.com/index.php
Proteome Discoverer	Thermo Fisher http://www.thermoscientific.com/wps/portal/ts/products/detail?productId=11961811
Scaffold 3 Q+	Proteome Software http://www.proteomesoftware.com/QPlus/ScaffoldQ+.html
Skyline	MacCoss Lab Software https://skyline.gs.washington.edu/labkey

Table 2: Major Mass Spectrometry Data Analysis Software.

Software	Platform	LIMS	High Level Data Processing	Label & Label-Free Support	Cost	Data Ware-house	Instrument Specific
MaxQuant	Desktop	No	Quant.	Yes	-	No	Thermo
PepTracker	**Web & Desktop**	**Yes**	**External Tools, e.g. MaxQuant**	**Yes**	**-**	**Yes**	*
Progenesis	Desktop	No	Quant.	Label-Free Only	from £16,100	No	Various
ProteinScape	Desktop	Yes	Quant. (WARP-LC Module)	Yes	Yes	No	Bruker
PROTEOIQ	Desktop	No	Quant.	Yes	from $499 (limited to 40 files)	No	*
Proteome Discoverer	Desktop	No	Quant.	Yes	£10,000	No	Thermo
Scaffold 3 Q+	Desktop	No	Quant.	ITRAQ Labelled	$5,995	No	*
Skyline	Desktop	No	Quant.	Label-Free Only	-	No	Various

** Software does not deal with raw data, instead takes data from external search engines.*

Table 3: Comparison of Major Mass Spectrometry Data Analysis Software.

Table 3 shows that, besides PepTracker (the software developed during this research), all of the software are desktop based, not making use of a web platform. Furthermore, some of the software, namely Progenesis, Scaffold 3 Q+ and Skyline, limit themselves to work with data generated from label-free or labelled samples only. None of the other software incorporate a LIMS, apart from ProteinScape, which couples their LIMS closely with Bruker instruments. Another feature lacking in the software is the ability to warehouse data from multiple researchers and experiments, which would allow for further extensive analysis. In addition, commercial companies were behind the development of all software with the exception of MaxQuant, which often results in the development of software being displaced from the actual science carried out in research laboratories. PepTracker's unique focus on downstream analysis coupled with data warehousing sets it apart from the other software reviewed, this is evident by the built-in, centralised, data warehouse in PepTracker. Rather than duplicating

functionality already available in other software, PepTracker in fact makes use of MaxQuant for quantitation of mass spectrometry data in its workflow.

1.2.3 Experimental Data Workflow

Solutions are needed that can transparently take the data being generated by MS instruments and interpret it so that it may be stored in a manner that facilitates future analysis and processing. This is not a simple task due to the large variety of MS instrument platforms, such as ion traps, quadrupole/time-of-flight, ion cyclotron resonance and time-of-flight/time-of-flight, which have caused a number of different proprietary data formats and incompatible systems to be developed, adhering to their own rules and structures. Often tools that are developed for data processing are tied-in with instruments, such as the WARP-LC module developed by Bruker for use with data produced from Bruker instruments, making the task of comparable data analysis near impossible. The various file formats associated with the range of instrumentation include XCalibur/RAW (Thermo Electron), Analyst/WIFF (ABI and MDS Sciex), MassLynx/RAW (Waters) or BAF (Bruker). This tactic of coupling data formats to instruments ensures that researchers are restricted to using hardware and software from one instrument provider, hence benefitting companies. Due to the proprietary nature of such software, they are not developed or documented extensively because the companies creating them have a greater emphasis on the development of instrumentation, which is the core focus of their business model. Due to these reasons, writing software to be vendor neutral is very important. This issue has been tackled by third party software, such as RAW2msm and the Trans Proteomic Pipeline (TPP), which convert mass spectrometer files to readable file formats that can be passed into proteomic search engines for further processing. In order to carry out database searching, search engines, such as MASCOT (www.matrixscience.com) (Perkins et al., 1999), SEQUEST (Eng et al., 1994) or X!Tandem (Fenyo and Beavis, 2003), require proprietary data formats to be converted into common text file formats, such as mzXML or mzML, and, in some cases, to mgf (MASCOT), dta (SEQUEST) or pkl (ProteinLynx, Micromass) (Nesvizhskii and Aebersold, 2004).

Once search engines have the data in a readable format they have the capabilities to convert the mass spectral data into peak lists with associated intensity values. Using these peak lists the software then carry out the task of identification, using externally

available sequence databases, which involves mapping the mass spectra produced by an MS instrument to identified peptides and thereby map to genome-encoded proteins. They are also responsible for filtering out less reliable identifications. This is one of the most important phases of high-level data processing in a proteomics pipeline. After protein and peptide identifications have been made, quantification software, such as MsQuant (Schulze and Mann, 2004) and MaxQuant (Cox and Mann, 2008), can be used to extract meaningful ratios (for labelled protein/peptide scenarios) or intensity (for label-free scenarios) values from multiple mass spectra. This further processing is carried out in quantitative experiments to determine protein expression levels. The outcome of this upstream processing is a set of textual data that should be easier to handle. However, further complications arise, as only specific versions of Microsoft Excel, the choice tool for analysis by scientists, can open these large files and in some cases the files are too big to be opened in all versions of Excel and require either command line access, or a more basic application, such as Notepad.

Figure 5: Mass spectrometry data workflow.

1.2.4 Workflow Management Systems

As described above, the processing involved with proteomics data is a multi-step procedure (Kearney and Thibault, 2003, Baldwin, 2004). Due to the various tools that have been developed, each performing specialised tasks, there is an obvious need to create technologies that can connect data and tools. However, when trying to achieve this task, the issue of interoperability arises, whereby the appropriate standards and infrastructure are not available leading to inter-application communication problems (Neerincx and Leunissen, 2005). The lack of standard data formats and definitions has been a major problem for a long time. It has been suggested that the issues related to inter-application communication can be overcome through implementation of an automated linear pipeline (Domon and Aebersold, 2006). The Workflow Management Coalition (WfMC) describes this as a Workflow Management System (WMS) and provides the following definition: "[A WMS] defines, manages and executes workflows through the execution of software whose order of execution is driven by a computer representation of the workflow logic". A number of attempts have been made to

represent the daily work of scientists within an automated pipeline using a WMS. Some examples of these include:

- Trans-Proteomic Pipeline (TPP) (Keller et al., 2005)
- Taverna (Oinn et al., 2004)
- Biopipe (Hoon et al., 2003)
- BioWMS (Bartocci et al., 2007)
- Wildfire (Tang et al., 2005)
- Pegasys (Shah et al., 2004).

The generation of automated workflows can involve the creation of a central data repository, which is often a Relational DataBase Management System (RDBMS). Data from the multiple software packages in the pipeline feed into this central database. Often, the software fitting into the pipeline can also be customised. However, many of the pipelines mentioned above are restricted to working with specific software that have an accessible interface and new pieces of software that do not conform to the standards defined by the WMS can not be incorporated into the pipeline.

Another issue that arises in the creation of pipelines in life sciences research is transparency. Analysis of use cases from the BioMOBY (Wilkinson et al., 2003) project have shown that scientists often want to view intermediate results to understand why an automated pipeline produced certain results. This is not always apparent in workflows, which simply take an input and produce a set of results as a standalone output. Developers should thus maximise the traceability, reproducibility and a compositional nature within an automated workflow (Bartocci et al., 2007). Other challenges within existing workflows include the cost and effort involved in maintaining solutions. These issues are enhanced due to data and software changes that occur frequently in a rapidly advancing research field, such as proteomics.

1.2.5 Potential for Collective Data Analysis

When new techniques became available for large-scale protein identification, many people involved themselves in the production of large quantities of data that were poorly organised and quickly became impossible to manage. Developing a well-documented data store means that analysis tools can be built to process the data without being concerned with the disparities in data formats of the original

instruments or upstream processing software. The development of software, to analyse data, can allow scientists to make new discoveries that would otherwise require great effort. The tools created must produce accurate and reproducible results through a transparent analysis process (Nesvizhskii et al., 2007). An analysis of the results can be presented in various forms due to the variety of questions that biologists may want answered. Hence, it makes sense for analysis software to consist of separate components that can each retrieve and examine specific data in defined ways. Individual researchers, generating data, will currently carry out the downstream analysis of their quantitative proteomics data. Due to the lack of structure and organisation of quantitative data there has been very little work on developing applications to carry out automated analysis. Automated analysis could potentially reveal data contained within the spectra that may be missed on a first targeted investigation. This is a potential goldmine which presents an untapped avenue that could answer many biological questions that are simply too difficult to tackle in a manual effort.

1.2.6 Data Recording & Collection

In order to ensure successful analysis, it is also imperative to record the metadata describing the experiment and conditions used to generate the data being analysed. In addition, it is essential for researchers to share knowledge regarding the data processing parameters and quality (Domon and Aebersold, 2006). These metadata must be readily available in a common format, especially for comparative data studies. Having an agreed standard data definition allows researchers to directly compare related experiments and come to logical conclusions. Without this information, using pre-existing datasets can be questionable as their reliability cannot be guaranteed. Without actual metadata, scientists must make assumptions that can be incorrect and lead to false discoveries. Hence, the inclusion of metadata, both annotation and protocol information, must be included within standards relating to high throughput proteomics data. Published guidelines (Bradshaw, 2005, Bradshaw et al., 2006) have provided direction on what this metadata should consist of. These guidelines specifically mention the inclusion of the parameters used when running software that carries out database searching.

Research has shown that traditional approaches to metadata collation and management, i.e. laboratory notebooks (lab books), are no longer meeting the demands of multi-user, multi-tasking labs (Piggee, 2008). The problems with lab books can be summarized into two main features: they are paper based and difficult to search. Paper based solutions require greater space to store and are therefore often locked away in locations that are not easily accessible. Furthermore, finding the correct lab book can be impossible and attempting to read the handwriting in the book, even more difficult. Problems also arise with issues of security and in keeping reliable and archived copies. In contrast, computers are ideal tools for providing quick access to large quantities of data. Automation of such activities has proven to increase productivity as well as aid in the advancement of research (Smallmon and Ganjei, 2004).

By recording the various steps of an end-to-end experiment and analysis computationally, it is possible to collect extensive metadata that can play a pivotal role when revisiting experiments and analyses carried out in the past or when attempting to use previous data in new analyses. This includes documenting where the original raw files are stored during the MS instrument analysis so that the original data can be re-processed using newer versions of software. Scientists often carry out this type of data management locally, on their own workstation, which makes tracking of samples from the mass spectrometer to raw mass spectra and finally to the protein and peptide identifications impossible. Further problems arise when the researcher leaves the laboratory, taking a whole host of data with them and rendering the re-analysis of their data impossible.

1.2.7 Data Dissemination

On a wider scale, further technological improvements, organisation of international proteomics projects and open access to results are needed for proteomics to fulfil its potential (Tyers and Mann, 2003). Modern life sciences research laboratories are no longer environments in which scientists work alone, they have progressed to carrying out work, often with global ties, that is more quantitative and comparative in nature (Domon and Aebersold, 2006). The lack of common data formats and standards makes it near impossible for researchers, from multiple laboratories, to compare their datasets generated using various instruments and platforms. In fact, data

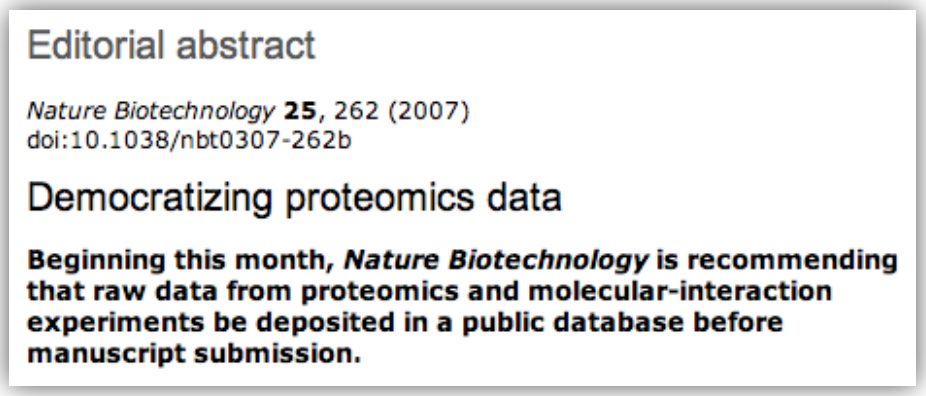
dissemination has been described as the weakest area in proteomics research (Prince et al., 2004, Rohlff, 2004). The reasons for this are two-fold, lack of technology to accomplish dissemination of data and the attitudes of researchers. Attitudes of researchers play a significant role due to the competitive nature of many scientific fields, and issues of both psychology and peer-review based funding mechanisms, whereby free scientific communication of results is often not encouraged to provide the originators of the data with a perceived competitive advantage and possibly help them to obtain further funding.

Apart from the sociological reasons, there are also many logistical reasons, which make it difficult to share data. These reasons relate to local data management, standardisation and availability of public repositories (Martens, 2006). In cases where researchers do make their data available, it is commonly in the form of published lists of proteins in a PDF. This makes the task of comparing data extremely tedious and difficult for any substantial datasets. Not only is extracting data from a PDF problematic but also the data are formatted differently from one PDF to another, for example different accession systems may be used in datasets produced by different users. Furthermore, in the field of proteomics, data age quickly, as illustrated by the frequently changing protein accession numbers, resulting in published data becoming out-of-date and inaccessible almost as soon as it is produced. Hence, there is a need to archive these data in accessible repositories that can handle changing data over time to promote sharing of data. This requires the use of common formats and standards to ensure the data are easily readable and can be shared. Currently researchers are unable to share datasets unless they are running samples on the same MS machine and then using the same database-searching program.

It is imperative that data dissemination issues are overcome to create a transparent process of data communication that can benefit the global research community. The benefits this would bring include consistent and more efficient assessment by reviewers, additional citations and most importantly, the potential to generate new results that will advance the field. In recent years, one of the efforts to try and aid with sharing large datasets is Tranche (https://trancheproject.org). Tranche consists of a network of computers, to which users can upload files and download files. Files can be of any type, size and with encryption if required. However, Tranche acts only as a

file storage and dissemination system, without focus on the content of files. Without knowing which file you are looking for and the meaning of the data defined in the file it is impossible to make use of the data – Tranche is designed only as an ideal solution for sharing of large files. Hence data management solutions are required to hold data in a searchable manner and provide a suitable format for sharing these data.

1.2.8 Public Data Repositories


Editorial abstract

Nature Biotechnology **25**, 262 (2007)
doi:10.1038/nbt0307-262b

Democratizing proteomics data

Beginning this month, *Nature Biotechnology* is recommending that raw data from proteomics and molecular-interaction experiments be deposited in a public database before manuscript submission.

Figure 6: Abstract from Nature Biotechnology.

Based on the continuing realisation that data should be shared and openly available in the life sciences community, journals are now recommending that data be deposited in public repositories that are publicly available (see Figure 6). Repositories have been created to capture data and fill the gap that has become evident. Public repositories, such as the PeptideAtlas (Institute for Systems Biology) (Desiere et al., 2005), Global Proteome Machine Database (GPMDB) (Beavis Informatics) (Craig et al., 2004) and the PRoteomics IDEntifications (PRIDE) Database (European Bioinformatics Institute) (Jones et al., 2006), have all developed due to the needs of the constantly evolving experiments and datasets. Even though these repositories have been extended and expanded to include original spectra, links to associated proteins and genomics data, they are still unable to handle quantitative proteomics data. In addition, these repositories are mainly aimed at capturing published data as opposed to the immediate data that researchers are working with. This raises the need for the creation of local repositories for use within laboratories.

Public repositories often suffer from problems related to standards. If public repositories are following specific standards then it becomes difficult for researchers to understand and use the data for their own purposes. Furthermore, public

repositories often accept data of varying quality from a wide range of authors from different laboratories. It cannot be guaranteed that all of these sources will be producing data of similar quality and reliability, as would be preferred or required when documenting new discoveries. Hence, this leads to the need for privately owned repositories, managed by laboratories, that link into external repositories for supplementary data. These types of repositories must also be well maintained and extendible to allow them to cope with the fast paced development of the proteomics field and provide data to researchers in a readable format that is easily understandable and usable.

1.2.9 Data Standards

To make data accessible, laboratories must decide how to structure data when they are making it available to others. However, either creating their own standards or using specialised protocols often adds to the problem, as other researchers will not be motivated or have the time or resources to incorporate data into an analysis if these data differ in format to their own. Thus, researchers must design and develop common standards that are universally readable. This situation is being tackled by organisations such as the international HUman Proteome Organisation (HUPO) (Hanash and Celis, 2002). In 2001, HUPO launched a project to involve the entire international community in discussions to enable reaching a consensus over standards. In order to aid with tackling standardisation issues, HUPO have developed the Proteomics Standards Initiative (PSI) (Orchard et al., 2005). Standards and guidelines are already under development in the form of the Minimal Information About a Proteomics Experiment (MIAPE), the format for mass spectrometry data (mzData) as well as the standard for analysis data (analysisXML). Further published standards (McDonald et al., 2004, Pedrioli et al., 2004) and guidelines are available for those aiming to set standards (Carr et al., 2004, Bradshaw, 2005, Wilkins et al., 2006). In all of these efforts, one of the most important issues highlighted is the reasoning that standards imply that everyone conforms to a specific format, hence all attempts at creating proteomics data standards must converge at some point, otherwise researchers shall be left with a number of so-called standards to choose from. When creating standards it is imperative to involve all relevant parties in the discussions, including manufacturers of instruments, the researchers carrying out ‘wet’ laboratory experiments and the software developers creating the processing and analysis software. It is encouraging to

know that two of the main proteomics standards (mzXML, developed by the Institute of Systems Biology in Seattle and mzData) are being merged into one new standard mzML. MzML is a data format for the exchange and storage of mass spectrometer files. The generation of these new standards has been an arduous process as there are many sources and types of data generation in the proteomics field. At the time of writing there are no agreed standards, draft or final, for quantitative proteomics data generated during analysis of MS data.

Creating standards for data in the proteomics field is far from simple. The complexity of proteomics leads to many issues when setting up new software, much of this centring on the data being stored. An example of this would be protein annotation. Despite the major advances made in documentation of metadata regarding proteins, there are still many disagreements. One of the dominant issues surrounding protein annotation is agreement on accession numbers. This stems from the varying definitions of a protein in the highly dynamic environment of the cell. For example, a researcher looking at a specific pathway in the cell will identify and publish a protein under a new name and hence a new protein will be defined. However, at the same time the set of amino acids making up the protein may be identified by another researcher looking at the structure of a protein complex, hence the second researcher would also rightly define this as an entirely new protein. These two instances that are resultant from this scenario are in fact the same protein. As time goes on, with further investigation, understanding of the two documented proteins is likely to be merged to one novel protein. However, this type of scenario means that any database system must be able to deal with this constantly changing data description and redundancy. To tackle this issue, various data sources have developed their own accession numbering system, which users are then forced to translate between. A further complication of these varied accession systems arises due to the redundancy in the accession numbers making them highly unstable, likely to be updated and occasionally deleted. As well as changing continuously, multiple identifiers within one source database may actually refer to a single protein. This form of redundancy is very common and expected in data sources related to proteomics.

Without a common accession number for proteins, software developers have the challenging task of continuously converting between accession numbers to carry out

comparisons and retrieve information from different sources. To aid in this task, a number of protein mapping services have been created to simplify the task of finding proteins in multiple data sources. The Protein Identifier Cross-Referencing (PICR) Service (Cote et al., 2007) is one such tool. It provides an interactive web interface and a Simple Object Access Protocol (SOAP) web service to translate between accession systems. It has the capabilities to access identifiers from over 60 source databases and is built around the UniProt Archive (UniParc), a non-redundant source of protein sequences.

The life sciences community have to address the issues described above in relation to data acquisition and integration, processing, analysis and dissemination to make progress in the field. The recent developments in instrumentation and scientific protocols have led to data generation on a scale previously unknown to the field of proteomics and almost unique in life sciences, with gene sequencing as one of the only other examples of large-scale, singular data generation endeavours. The problems described make it evident that advancements in science have to be supported by adequate developments in the field of computing. There are many new and exciting discoveries to be made through cross-disciplinary work that brings together science and computing in a usable fashion that can enhance knowledge and understanding in both fields. With the right support and development, the way in which data management and analysis is tackled in laboratories can be revolutionised to support novel avenues of analysis that were previously impossible. This opens up the possibility to formulate and evaluate biological hypotheses that could lead to new biological discoveries.

1.3 Data Visualisation Approaches

1.3.1 Origins of Data Visualisation

Data visualisation is an interdisciplinary subject area being investigated and applied to heterogeneous data sources across many applications. Even though data visualisation has mainly developed over the past 30 years, its origins come from 2nd Century Egypt where astronomical measurements were first recorded in table format (see Figure 7) (Few, 2007). Current age data visualisation unites two sub-areas, namely scientific visualisation and information visualisation (Post et al., 2003). Card et al. describe the difference between scientific visualization and information visualization as primarily

based on the difference in input data. Scientific visualisation is applied to physical data that often has a natural spatial mapping whereas information visualisation techniques apply to abstract data without inherent spatial mappings (Card et al., 1999). However, this distinction between the fields has been challenged by interest in visualisation by the bioinformatics community for genomic data (Rhyne, 2003). It has been rightly identified that “information visualization is not unscientific, and scientific visualization is not uninformative” (Munzner, 2000).

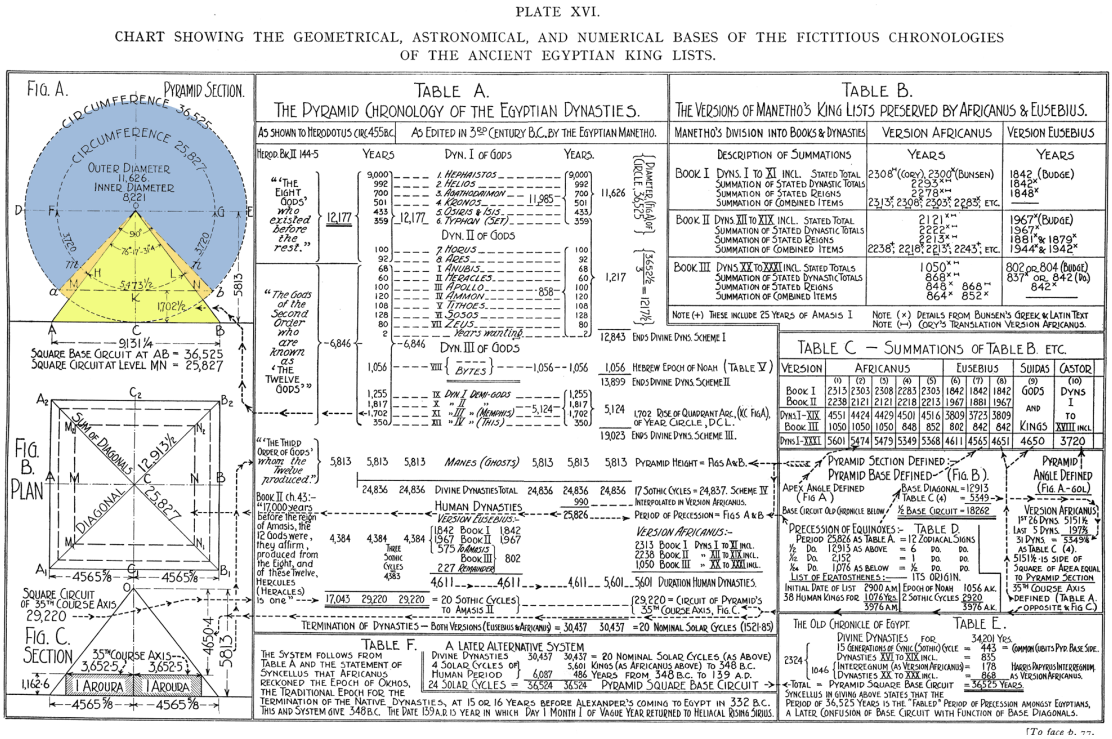


Figure 7: Egyptian astronomical measurement table.

In 1983 Edward Tufte, described as the “Leonardo da Vinci of Data”, by The New York Times, noted that “At their best, graphics are instruments for reasoning about quantitative information. Often the most effective way to describe, explore, and summarize a set of numbers -- even a very large set -- is to look at pictures of those numbers. Furthermore, of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful.” (Tufte, 1983). In 1987, a report was compiled by the National Science Foundation (McCormick et al., 1987) describing visualisation as "the use of computer graphics to create visual images which aid in understanding of complex, often massive numerical representation of scientific concepts or results". Neither of these prominent publications stated a difference between scientific and information visualisation (Rhyne, 2003). Both of these sub-domains aim to use visual

representation to facilitate knowledge discovery by taking advantage of the human ability to spot patterns, trends and outliers.

1.3.2 Types of Data Visualisations

Visualisations centre on displaying data using coordinate systems, colours, lines, points, shading, numbers, symbols and words (Tufte, 1983). The various reasons for data visualisation have prompted a number of different visualisation components to be developed. One class of visualisations is graphs, tables and diagrams, which collectively fall under the category of charts. Even though the traditional table of data is overlooked as a visualisation, the way in which data are visually arranged into rows and columns make it one of the most popular and effective ways of presenting data. Graphs, with their organisation of data using two-dimensional coordinates, came much later and researchers now commonly use charts, such as pie and bar, to convey information.

Data features coupled with the way data is to be used and the context play a role in how a visualisation is designed. The type of data involved will influence how it is presented and mined. Data can vary from being:

- Hierarchical: organisation of data into several levels where each level is linked to the one above and below it, for example cells can be classified into species, organism, organ, tissue and so on,
- Spatial: where data is separated by location, for example data at the cellular level can be collated for different cellular compartments such as cytoplasm, nucleus, nucleolus etc. and/or
- Multi-dimensional: data is defined by several dimensions, for example an experiment can be described by many dimensions, for example cell type, organism, treatment, date etc.

There are three main purposes for visualisations: data exploration, hypothesis confirmation and visual presentation (Yeh, 2006). John Tukey defined data exploration as exploring data where the researcher does not know what they are looking for (Tukey, 1977). It has been found that spatial visualisations are the most effective in this case as they allow unconstrained searching and undirected navigation to uncover trends and anomalies. This is often carried out as a first stage in analysis and can lead

to hypothesis generation. When a researcher would like to confirm a hypothesis they believe to be true they need to be able to see relationships in data clearly. For this task, traditional 2D scatter plot representations of data can be very informative. For visual presentation of data, the most important consideration should be the ease-of-perception for unfamiliar viewers.

In recent years the web has become a viable medium for visualisation, which has prompted design of new components that have built-in interaction, animation and graphics. The popularity of web-based visualisation has been further enhanced through larger companies becoming involved in the development of libraries of visualisations. Google Finance is an example of a web-based tool that has developed a standard for interactive timeline charts and Google Maps has defined how interactive maps should be used. Furthermore, Google have created Application Programming Interfaces (APIs) such as the Google Chart API (see Figure 8) and the Google Visualisation API to allow software developers to quickly and easily incorporate Google style charts into their own web applications. IBM is another company delving into the world of visualisation through their offering: Many Eyes, which is an internet based tool allowing users to upload their data and view it in a serious of pre-defined visualisations that are interactive. The Many Eyes visualisations can then be shared through embedding into web pages etc.

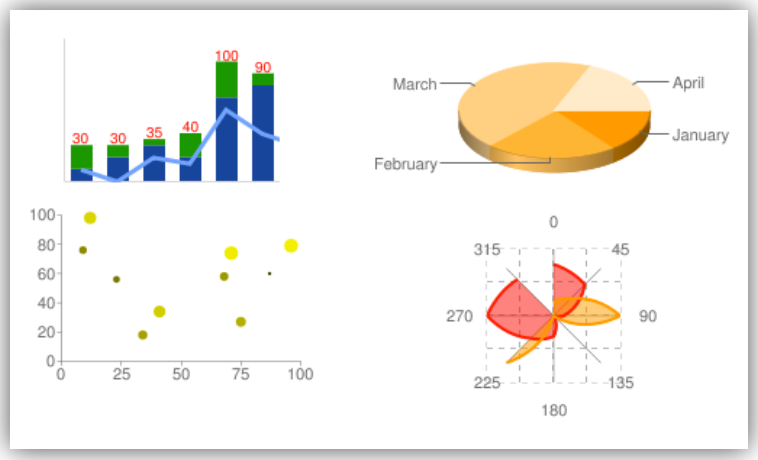


Figure 8: Examples of charts available using the Google Chart API.

1.3.3 Potential for Data Visualisation in Life Sciences

The main driving force behind data visualisation is the desire to make it easier to reveal important information in complex data and accessing the ability to interact with

data in novel ways. Furthermore, visualisations are an invaluable tool for communication of results in research environments. Visualisations provide an alternative interface for working with data that make use of human perception abilities, for example researchers can gain insight and spot visual anomalies more intuitively in graphical representations compared with lists of numbers.

Within life sciences many research laboratories follow the traditional route of analysing data, which primarily revolves around an 'Excel culture', whereby researchers extract data from multiple software and load them into spreadsheets and perform their own calculations, making use of in-built Excel functions and charts. This raises the possibility of researchers failing to identify interesting discoveries, especially in cases where a researcher has limited experience of analysing and manipulating the type of data they are currently handling. By implementing standard visualisations that are transparent, researchers could benefit both in scientific productivity and raise the potential for major science breakthroughs, described as being on par with supercomputers (McCormick et al., 1987).

The complexity of proteomics data makes it more difficult to understand if only viewed as data in tabular format. In comparison to tabular data the human brain is more susceptible to intuitively recognising patterns in visual displays. The best pattern matching algorithm available, as well as the final decision making authority, is still the mind. Researchers rely on their experience and knowledge to draw reliable and well-informed conclusions, which is difficult if not impossible to capture in a computer system. Hence, using visualisations to complement automated data mining and knowledge discovery can aid to uncover useful knowledge.

"Information visualization is becoming more than a set of tools and technologies and techniques to understand large datasets. It is emerging as a medium in its own right, with a wide range of expressive potential." (Eric Rodenbeck, Founder & Creative Director of Stamen Design).

Data visualisation provides researchers with another tool in their analysis of often very complex and increasingly large volumes of data. This was realised in the National Science Foundation report which described the problem and potential role of visualisations in science stating "Users from industry, universities, medicine and

government are largely unable to comprehend or influence the 'fire hoses' of data, produced by contemporary sources such as supercomputers and satellites, because of inadequate Visualization in Scientific Computing tools". Shared data visualizations can further help with collaboration between researchers in science.

Visualisations have extended in recent years, from basic 2D and 3D environments to dynamic multi-dimensional environments. Multi-dimensional data, which describe data entities with more than three attributes, is becoming more popular in data visualisation. There are some excellent examples of commercial multi-dimensional software in industry, including an in memory multi-dimensional analytics tool developed by Spotfire (http://spotfire.tibco.com/). The Spotfire DecisionSite is a decision-making software tool that helps users identify relationships and patterns in data through interactive, visual approaches. With the generation of more complex data and emergence of data warehousing and data mining, the application of visualisation to high dimensionality datasets has become more relevant and required.

It should not be underestimated how useful static graphs, both on paper and in electronic form, are in communicating a great deal of information. However, the layer of interaction between researcher and data also plays a significant role when researchers are at the exploration and analytics phase. This has prompted further research and developments in the Human Computer Interaction (HCI) field, which looks at the interaction between user and graphical user interface. The development of interactive interfaces, such as the point-and-click mouse interface, the what-you-see-is-what-you-get (WYSIWYG) interface, drag-and-drop interfaces and hierarchical file browsers, have pushed visualisations to evolve to provide increased interactivity in intuitive ways. By providing a more immersive environment with suitable interactivity, researchers have the ability to quickly assimilate and interact with their data and generate hypotheses.

1.3.4 Problems with Data Visualisations

Data visualisation is often greatly under estimated due to lack of understanding. Current visualisation tools in life sciences suffer from problems making them less efficient in meeting their full potential in the scientific analysis pathway. The lack of real-time interactivity with simple charts can be frustrating for researchers. On the other hand, making visualisations very complex can also make them more of a

hindrance. In these situations, researchers will opt to discard and fail to adopt these visualisations into their standard analysis workflow. This can lead to researchers being disillusioned as to the full advantage of effective visualisations in exploration, analysis and communication of their data. A successful visualisation should initially translate understanding of data to a researcher through immediate perception of outliers, clusters and/or trends. Follow up interactivity with the visualisation should provide researchers with the ability to delve into their data. Superfluous functionality that simply makes visualisations look good should not interfere with exploration of data for useful research.

It is imperative to keep sight of the primary goal of data visualisations in life sciences research, which is biological insight. Often developers will focus on the development of impressive visualisations and fail to meet the aims of the scientific discovery process (Bethel et al., 2009). It is also important to consider how users will interact with data and then create an appropriate human computer interface. In order to ensure this is implemented well, interactions with visualisations should be tested with researchers to ensure they add benefit to analysis, rather than act as a distraction or hindrance. However, objectively evaluating visualisations remains a fundamental issue that is challenging to address due to the difficulty in defining objective criteria (Johnson, 2004). The phrase: “overview first, zoom and filter, then details-on-demand” from The Visual Information-Seeking Mantra (Shneiderman, 1996), is a basic principle summarising an excellent framework for designing visual data applications. Scientific analysis centres around “What-if” type of analysis, which makes the interactive nature of visualisations a valuable tool for investigation of data to answer questions efficiently and effectively (Johnson, 2004).

When creating visualisations it is also important to consider screen sizes on which the end-visualisations will be viewed. It is key for researchers to have the ability to view a complete visualisation on screen when attempting to determine trends and patterns effectively. If a researcher is unable to view all data points then they are unable to perceive the potential relationships that could lead to interesting discoveries.

Hans Rosling, who is a world health expert and data visionary, believes that making information more accessible has the potential to change the quality of the information itself. Rosling has demonstrated his belief in a series of TED talks using specially

designed software, GapMinder (see Figure 9), to show how statistics and data can be made more accessible and lively. His talks have given him global prominence as a visionary in the field.

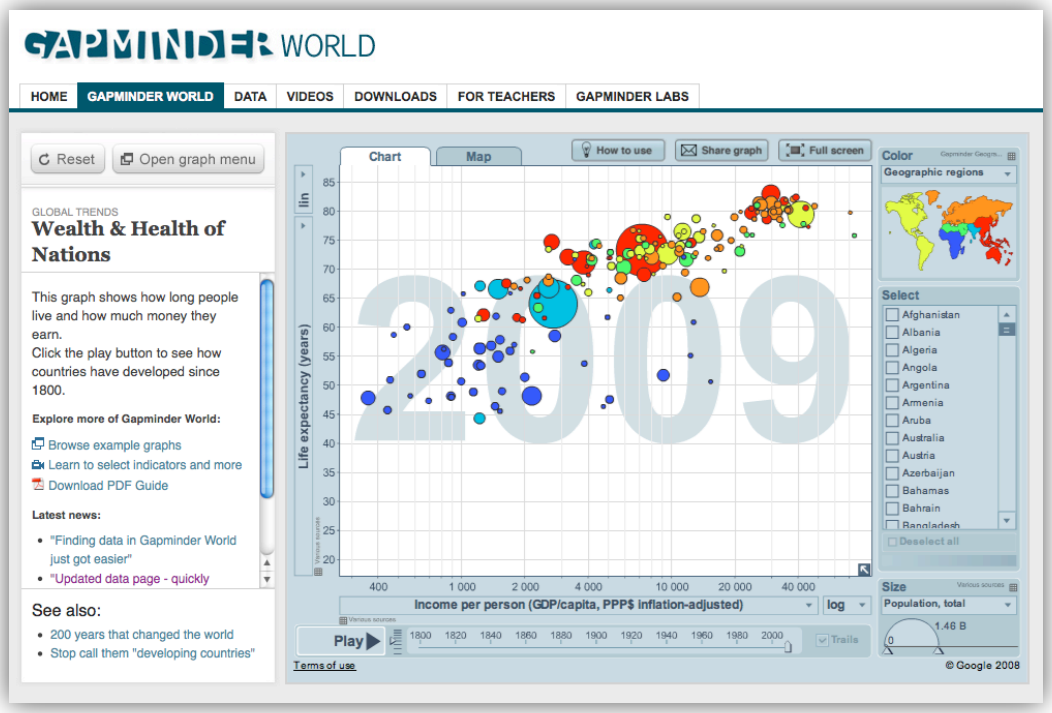


Figure 9: Screenshot of GapMinder (http://www.gapminder.org).

In order to make accurate hypotheses and publish data in peer-reviewed journals it is important to ensure data are of a high standard, however many visualisations do not take this into account and do not let researchers manipulate the data being displayed. Data quality issues are ubiquitous within scientific measurements. It can be difficult to comprehend their impact when viewing quality parameters as numbers in a list. However, by visually seeing the data quality measurements researchers are less likely to base hypotheses founded on irrelevant, incomplete or questionable data. This was realised early on in the visualisation world with the National Science Foundation report on visualisations stating “Scientists need an alternative to numbers . . . The ability of scientists to visualize complex computations and simulations is absolutely essential to insure the integrity of analyses, to provoke insights and to communicate those insights with others.” (McCormick et al., 1987). This is very much true in the proteomics field, as having a good understanding of data can lead to more reliable hypotheses and follow up studies. One example of visualising data quality is through dynamic filtering of values on a graph.

Often developers of visualisations fail to spend adequate time considering and understanding the underlying scientific data. This can be achieved by working directly with researchers generating the data. It has been reported that there is no substitute to working side-by-side with end users (Brooks, 1996, Johnson, 2004). It was predicted early on, in the National Science Foundation's report, that visualisations would be the mechanism that would bring researchers into the computing loop (McCormick et al., 1987). The report insisted that in order to develop successful visualisations it would be important to foster interactions between researchers and visualisation experts. Interdisciplinary teams are highly recommended as researchers have a unique insight of the important factors in the underlying scientific data they have generated and visualisation experts have the skills to implement high quality visual tools.

1.3.5 Web-Based Visualisation Technologies

With the web increasingly being used for data accumulation, analysis and dissemination, a number of technologies have arisen to allow programmers to create rich interactive applications (RIAs). These types of applications benefit from a state-based client environment, whereby the researcher does not have to reload pages or move between many pages. Technologies such as SilverLight (Microsoft), Flex (Adobe Systems) and Google Web Toolkit (GWT) have the ability to send and receive data from servers dynamically, without the need to reload the browser. Furthermore, these technologies come with a collection of rich libraries that can be used to build visualisations. A report by Gartner predicted that "Interactive visualisation will be quickly accepted during the next two years as a common front end to analytical application, driven by the ubiquity of rich Internet applications." (Schlegel, 2008).

The main difference between the three popular RIA's is their mode of development, with GWT compiling Java code to JavaScript, Silverlight compiling XAML to a XAP file that runs in the Silverlight plug-in and Adobe Flex compiling a combination of MZML and Actionscript to a Flash swf file. Adobe Flex has more components, both built in and available from the open source community, as compared to GWT and Silverlight. Another major benefit of Adobe Flex is its partner product, Adobe Air, which allows deployment of Adobe Flex applications on both web and desktop platforms. Compared with GWT, Adobe Flex also has the upper hand with regards to its multimedia user interfaces, with Flash Player providing a much more enhanced user experience

compared to a web browser. Silverlight requires users to download and install a custom Silverlight plug-in to run their content.

In recent years developments such as HTML5 are emerging as alternatives to interactive, flash based web solutions. HTML5 implements a canvas element and Scalable Vector Graphics (SVG), providing interactivity within a supporting web browser. In addition, Javascript libraries like WebGL (Web Graphics Library) can make use of the HTML5 canvas element to provide an API for rendering interactive 3D graphics.

However, unlike HTML5, tried and tested RIA technologies, such as Adobe Flex, work in a well-known and predictable run-time environment, i.e. the Flash Player. Adobe Flex has good performance, testing tools and internationalisation support. More recently, Adobe Flex has also been donated by Adobe to the open source Apache Foundation, which opens up the potential for the technology to grow with many developers able to contribute directly to its rich libraries. Furthermore there are many convenient tools for the development of Adobe Flex applications, including integrated development environments (IDEs), compilers, debuggers and profilers. These tools make the development and testing of web based applications much simpler than attempting to accomplish the same functionality with a combination of HTML5 and several other technologies, including AJAX, JavaScript, CSS and the XMLHttpRequest object. The use of several separate technologies results in the need for additional developer time and extensive, in depth testing to ensure the additional interfaces between the technologies work appropriately. Often, this added complexity involved in developing HTML5 applications results in a much lower functional specification for most HTML5-based web applications, due to the extensive developer time required for testing under several browsers.

1.4 Software Approaches for Biological Data Analysis

1.4.1 Super-Experiment Data Analysis

As mentioned in section *1.2.5 Potential for Collective Data Analysis*, typical analysis of data in proteomics is carried out manually by researchers on a single experiment, single dataset level. However, collection of datasets into a single data repository provides an ideal target data source that is large enough to be used as a baseline for

identifying global patterns and trends. Through the use of Knowledge Discovery (KD) techniques it is possible to provide automated analysis to extract implicit, unknown and potentially useful information from a target collection of datasets. These techniques can only uncover patterns and trends that exist in the data already, hence it is reliant on a data environment containing a large and continually expanding collection of consistently annotated MS datasets that are normalised and implemented in an n-dimensional database for classification, visualisation, probabilistic and statistical analysis approaches.

With this approach datasets could be linked and studied together to generate new hypotheses spanning broader questions. This approach has been termed as 'super-experiment' analysis. A super-experiment can be understood as analysis involving multiple independent datasets in which each dataset provides value to the analysis of every other dataset in the collection. By integrating the datasets in such a way, novel research can be carried out to identify patterns or trends that cannot be elicited from a single dataset. This analysis does not necessarily have to come from multiple related datasets, generated to answer specific questions, in fact it is envisioned that by analysing unrelated datasets it will be possible to extract further knowledge that was not hypothesised. In this way, each dataset has the potential of informing the analysis of existing and future datasets. This novel pathway of analysis necessitates tools that can collate a large number of organised and well-documented experiments that can then provide the opportunity to carry out super-experiments to extend into new biological analyses.

1.4.2 Applying Business Intelligence for Super-Experiment Analysis

The world of proteomics research is very knowledge centric. In order to make good research decisions and plans it is imperative for researchers to manage the data they are generating and maximise the information gleaned from the data. This is not dissimilar to the business world, which has been encountering and attempting to solve the same sorts of problems. Within the business world much research and effort has gone into the development of a field known as Business Intelligence (BI).

The term Business Intelligence was first used by an IBM researcher Hans Peter Luhn (Luhn, 1958). Luhn defined intelligence as "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired

goal". The field of Business Intelligence is primarily concerned with the development of tools and techniques to aid with the identification, extraction and analysis of business data. The aim of the computational techniques is to help businesses use their data, such as sales revenues, to support good decision-making. BI applications perform various tasks such as data integration, data quality, data warehousing, master data management, text and content analytics.

The role of BI in various companies usually involves the conversion of customer related data to information, which is then interpreted by analysts to produce knowledge that can be used to action improvements in the company. Even though BI strategies have been developed in the commercial arena and primarily used by businesses, there is no reason why these strategies could not be applied to data with different meaning in other fields, such as research. The types of processes involved in BI could also apply in the life sciences field where researchers want to convert their biological data into information that can then reveal useful knowledge. This knowledge could then be fed back into the experimental workflow, either through follow up experiments, or by improving existing experiments. Hence, it is hypothesised that BI could have a positive impact in the field of life sciences.

This has been demonstrated previously in an academic study carried out to analyse historical science data, which has enhanced understanding of how Darwin developed the theory of Evolution by natural selection (Kohn et al., 2005). Previously, this approach has not been used in an academic proteomics laboratory. It is hypothesised that wider application of these techniques will be of great utility, not only for academic proteomics research, but also for other research areas involving the collection and mining of very large datasets, as is now common in biomedical science.

BI deals incredibly well with the efficient analysis of large datasets. In particular BI principles have been designed with the aid of data warehouses that often contain data collated from many sources. If proteomics datasets could be arranged in a data warehouse, the use of BI could be made more straightforward. The core concept of BI revolves around understanding and modelling data in an appropriate format that makes analysis easier and more intuitive for end-users. BI technology is designed for rapid interactive response and works particularly well for train-of-thought analysis, whereby response times from queries are rapid enough (one to two seconds or less) to

allow a user to follow a sequence of ideas where each answer can prompt another question. The advantages of rapid response times on productivity have been well understood for many years (Lambert, 1984). BI techniques facilitate the analysis of complex data and are essentially discipline agnostic.

The BI method of analysing data includes OnLine Analytical Processing (OLAP). OLAP works alongside a data warehouse, which can be structured using a relational or multidimensional structure. The data warehouse is a vital component that must contain the required data in a consistent format. The data warehouse can be populated from various source systems to provide a comprehensive coverage of data needed to answer the questions of researchers. Data in life sciences is spread across many databases, hosted by institutions around the world that are specialists in their own areas. However, researchers would greatly benefit from being able to access differing data from one location, for example localisation, post translational modifications (PTMs) and domain information. Furthermore, researchers would ideally like to collate this information with their own datasets to draw conclusions. Searches to find these data are very time consuming. Some of the global databases available provide web access to data (O'Donovan et al., 2002), which can be used to make local copies of databases. To load data into the warehouse, it must go through an extract, transform and load (ETL) process to catch any variations in data schemas and data values. Once all data have been loaded into the data warehouse, OLAP can be used to transform the data into an OLAP cube, which is a multidimensional structure for querying and analysis.

1.4.3 Relational versus Multi-Dimensional Databases

A database is an electronic data store defined by a data dictionary, describing the fields and the various parameters associated with each field, such as data type and any constraints. Databases can be designed in a number of ways with two of the most popular being a relational and dimensional structure. A relational database models data by analysing the relationships between different data entities and defining those relationships. For example proteins are related to peptides as one protein is made up of multiple peptides. In comparison, a multidimensional database structure views data as a series of measures, typically values that are of interest, such as ion intensity of

peptides, and dimensions, i.e. the parameters used to extract specific data, such as cell type or date etc.

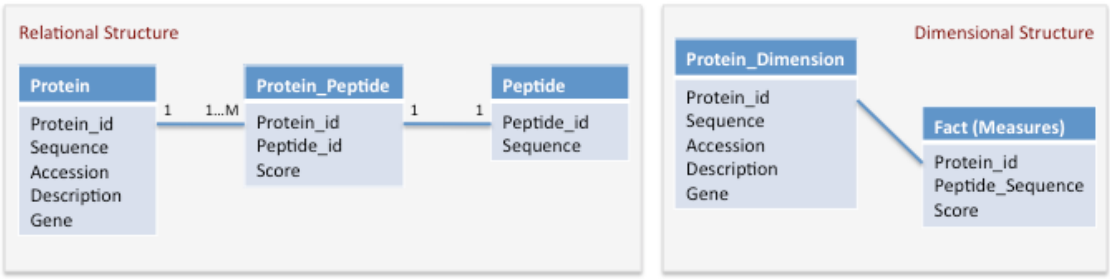


Figure 10: Relational versus dimensional database structure.

A relational database is inherently designed to handle transactional data. CReate, Update and Delete (CRUD) operations are handled very well by relational structures. Structured Query Language (SQL), a programming language, can be used to construct queries to extract data from tables in the database. There are many commercial relational database providers, such as Teradata, Microsoft, Oracle and IBM, as well as open source solutions, such as MySQL, SQLite and PostgresSQL. Within a relational engine data are not structured particularly well for analytics, as data can be spread across many tables, which have to be joined together before having the ability to extract the data of interest.

In comparison a multidimensional structure supports analytics in a more intuitive manner. A multidimensional data model can alleviate problems inherent in a relational database by making it easier to select, navigate and explore data. It is also able to provide increased query performance in comparison to a relational database, due to the way that it holds pre-aggregated data. Almost all query result times benefit from this type of pre-computation. However multi-dimensional databases are not particularly good for transactional, CRUD operations. MultiDimensional eXpressions (MDX) is a programming language allowing programmers to query multidimensional database structures. Commercial multi-dimensional database providers include Microsoft, Hyperion, Cognos and Oracle.

Chapter 2: Methodology

This thesis has resulted in the creation of an environment, PepTracker, for management and mining of cell biology data produced by scientists working in the specialised field of proteomics. The suite of tools within this software incorporates automated visualisation and analysis tools to handle quantitative data produced from proteomics studies. The creation of this software involved a heavy focus on involving users to drive the development aims and outcomes. Described in this section are the technologies and user-centred techniques that were applied to create this novel software in the expert proteomics field (section 2.1).

2.1 Software Development Approach

2.1.1 Developer Environment

During the creation of PepTracker, my location as the developer was thought to be very important. I was therefore embedded in the Lamond laboratory. This provided continuous contact with researchers and allowed me to gain insight into the needs and problems of the researchers from everyday contact. It also provided opportunity for passive ethnography to take place, for example lunch-time conversations led to the identification of issues that could be tackled easily via PepTracker, e.g. moving an external database (holding data about the components and solutions acquired by the laboratory) into the PepTracker system for easier management and linkage to experimental metadata. The researchers in the laboratory responded well to this approach with one proteomics researcher reporting "[Researcher] is constantly in touch and checking requirements of the individuals who use the data, is seated with them in a lab on a daily basis, and is acutely aware of our capability and limitations with regard to computing".

2.1.2 Ethnographic Observation

Ethnographic approaches were used to observe researcher tasks, carry out iterative evaluation and testing. These approaches focused attention on researchers and allowed for continuous refinement based on observations, for example, watching researchers enter experimental metadata helped identify a better database model to reflect the variations in experimental procedures. Furthermore, analysing researchers in their working environment is vital when creating visualisation tools that meet the needs of users.

2.1.3 New Ideas and Inspiration

Fostering an environment that encourages new ideas and inspiration is important to get researchers excited about the development. During the creation of PepTracker, there were many meetings that involved researchers describing complex user interface ideas that would improve their interaction with the charting components within PepTracker. Rather than refusing to consider these ideas because the code base would not allow for such developments, these requests were monitored. Feature requests included tasks like scrollable and zoomable charts. After realising the importance of these features to researchers, a new technology was incorporated into the architecture, i.e. Adobe Flex, to implement an interface that provided this increased charting interactivity. Furthermore, encouraging and allowing extensive 'thinking' about problems and data during researcher meetings, pushed the boundaries of the development to continue and explore new avenues, such as BI.

2.1.4 Domain Knowledge

Developing software for research laboratories is further complicated as research laboratories are at the forefront of discovery, hence their needs are often novel, specialised frequently changing. In order to understand how to deal with the changing needs of researchers of the PepTracker system, it was important to gain extensive domain knowledge. Within a scientific environment, this domain knowledge may be obtained by carrying out experiments. The tedious nature and steps involved in proteomics experiments was not fully understood until time was spent shadowing a biologist carrying out an experiment and also by taking part in tasks, such as pipetting etc., in the laboratory. From getting involved I was able to understand the data being processed and why there may be variability in the resulting data, e.g. from pipetting errors. In addition, observing researchers was imperative for the development of the PepTracker system, in order to fully understand and automate the workflow of a researcher studying proteomics.

2.1.5 Understanding Users

Involving highly intelligent researchers in software development can be challenging as researchers often find it difficult to translate their science into a set of requirements. The creation of PepTracker required an in-depth understanding of the problems faced by researchers in order to help them identify potential computational solutions. One

such problem involved developing a method of annotating proteins as contaminants in protein-protein interaction studies, by analysing very large sets of experiments. The overall solution to this problem was the use of BI principles to provide pre-aggregation of data for rapid response. Despite being new to the field of business intelligence, it was possible to implement the techniques by describing and helping researchers think of their problem in terms of 'measures', i.e. the values they are interested in, and 'dimensions', i.e. the fields on which they would like to query. Following on from this it was possible to create a Sun Model diagram, which describes a researcher's perception and understanding of their data, and then convert this into a physical implementation. This is the process used in the generation of the Protein Frequency Library (see Chapter 5: Multidimensional Analysis with IP Experiments).

Furthermore, by becoming involved in the biology of the project, it was possible to get in the role of a biology researcher and, therefore, consider the problems from a different point of view and have an increased motivation to solve the problem. It was also felt that this promoted the biologists to have a vested interest in making the development of the BI tool a success. This is evidenced in feedback received and by the willingness of researchers to take technical drawings, such as the Sun Model, and use these within presentations of their own projects.

2.1.6 Researcher Expectations

In order to manage the expectations from researchers, small iterations were delivered frequently. This allowed researchers to witness the development as it happened and to understand better the timescales involved in creating certain features. Also, the researchers were asked to prioritise requirements and hence help define a schedule based on a set of features they had requested, along with the projected timescales. This scheduling meant that researchers were more aware of what they could expect from the software and developer in a given timeframe.

2.1.7 Function and Form

It was also found that focusing on how a researcher achieved a task was just as important as what the task end-goal involved. Within PepTracker, identifying trends and patterns in data was a vital component. However, many of the meetings with users centred, not on what these trends and patters were, but rather on the process by which researchers would discover these insights and how tools could be developed

to facilitate this process. For example, a variety of chart features, such as overlaying groups of data, filtering of datasets and selecting ranges of data, were implemented as they helped improve user perception of the data being viewed and analysed.

2.1.8 Leadership

Software development can further be hampered due to a drive for productivity alone. Within a research area, it is important to understand the use of spending quality time simply talking about problems and data. For a project like PepTracker, where there was no existing similar software design, it was difficult for researchers to define what they wanted and for the developer to ask the right questions, hence emphasis had to be placed on the need to refine the questions until cohesive ideas could be obtained. Having leadership that allowed for the necessary scope and flexibility to explore new ideas, such as the use of BI, rather than placing pressure on output in terms of functionality, inspired innovation and novel interface design. In terms of the PepTracker system this came from the head of the Lamond Laboratory, Professor Angus Lamond, who is not a computer scientist and hence was interested by the biological significance of the work, yet was driven throughout the development to push for innovative ideas of tackling problems rather than measuring success by the number of papers released through the use of the software.

Chapter 3: Nucleolar Proteomics Database

3.1 Summary

An experimental data handling system has been created as an update to the previous Nucleolar Proteome Database (NOPdb3.0: http://www.lamondlab.com/NOPdb3.0/). This updated system is able to manage large datasets identified by multiple MS experiments and has been used to analyse highly purified preparations of human nucleoli from different cell lines. The newly created application includes a dynamic relational database, which is kept up to date by laboratory researchers. The data are further annotated with information from specific external sources on the web, including the IPI and Gene Ontology databases. In addition, an Application Programming Interface (API) provides external users with a portal to link into the nucleolar proteome database and hence gain access to continually updated results. From the initial ~700 human proteins identified in the previous iteration of the NOPdb, there are now over 50,000 identified peptides contained in over 4,500 human proteins from purified nucleoli, providing enhanced coverage of the nucleolar proteome.

Chapter 3 describes the NOPdb software, focusing first on the history of previous versions of the NOPdb (section 3.2), following with a description of NOPdb version 3.0 (section 3.3), its implementation (sections 3.4) and finally a discussion on the use of NOPdb3.0 (section 3.5).

3.2 Background

The nucleolus is a highly conserved nuclear organelle whose main function is to coordinate the synthesis and assembly of ribosome subunits (Boisvert et al., 2007). Previously, a Nucleolar Proteome Database (NOPdb2.0: http://www.lamondlab.com/NOPdb) was described that archived data on over 700 proteins that were identified by multiple mass spectrometry analyses from highly purified preparations of human nucleoli (Leung et al., 2006). Each protein entry was annotated with information about its corresponding gene, its domain structures and relevant protein homologues across species, as well as documenting its MS identification history, including all of the peptides sequenced by tandem MS/MS. Moreover, data showing the quantitative changes in the relative levels of approximately 500 nucleolar proteins were compared at different time points upon transcriptional inhibition (Andersen et al., 2005).

The data presented by the previous NOPdb, version 2.0, was held in a flat file database. Due to the aggregated nature of the data, results from individual experiments could not be extracted. The peptide data for a single protein were merged within this database rather than stored separately. The client interface to this database consisted of Perl CGI scripts. These scripts were able to extract the relevant data from the flat file database to create static html pages. After running the scripts, a page was created on the server for each protein. The html pages were then made available to the global community via the Internet. Each time data were updated in the flat files, the Perl scripts had to be run again in order to reproduce the static html pages. This process of having to reproduce the static html protein pages after each database update was highly inefficient and time consuming. A more efficient approach is to produce dynamic html pages upon user request. Furthermore, the capabilities of the NOPdb version 2.0 database were limited with respect to security, ease of use, accessibility, maintainability and expandability. For example, a number of security concerns arose regarding the Perl scripts, which proved very difficult to resolve due to limited documentation.

3.3 Nucleolar Proteome Database v 3.0

The new version of the NOPdb3.0 (http://www.lamondlab.com/NOPdb3.0/) consists of a unique, secure, extendable content management system, holding advanced nucleolar proteomics data. The created application includes a dynamic relational database, which is kept up to date by members of the Lamond Laboratory. It also allows the query of protein data hosted within the database by external users, either using the custom built graphical user interfaces, or by building custom web tools that access data via the custom API. In addition to the dynamic interfaces provided by the new content management system, the data included in the nucleolar proteome are also dynamically updated with proteins identified from several different cell lines, using various instruments, by members of the laboratory.

The new version of the Nucleolar Proteome Database (NOPdb3.0) archives all human nucleolar proteins identified to date by the Lamond Laboratory and their collaborators using MS analyses (Andersen et al., 2005, Leung et al., 2006, Boisvert et al., 2007). The current version 3.0 of the database is available at http://www.lamondlab.com/NOPdb3.0/ and is searchable either by protein name,

protein sequence, motif (Mulder et al., 2003, Bateman et al., 2004, Letunic et al., 2004), Gene Ontology (GO) (describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner) (Ashburner et al., 2000) terms or by setting the range of the predicted isoelectric point and/or molecular weight (see Figure 11).

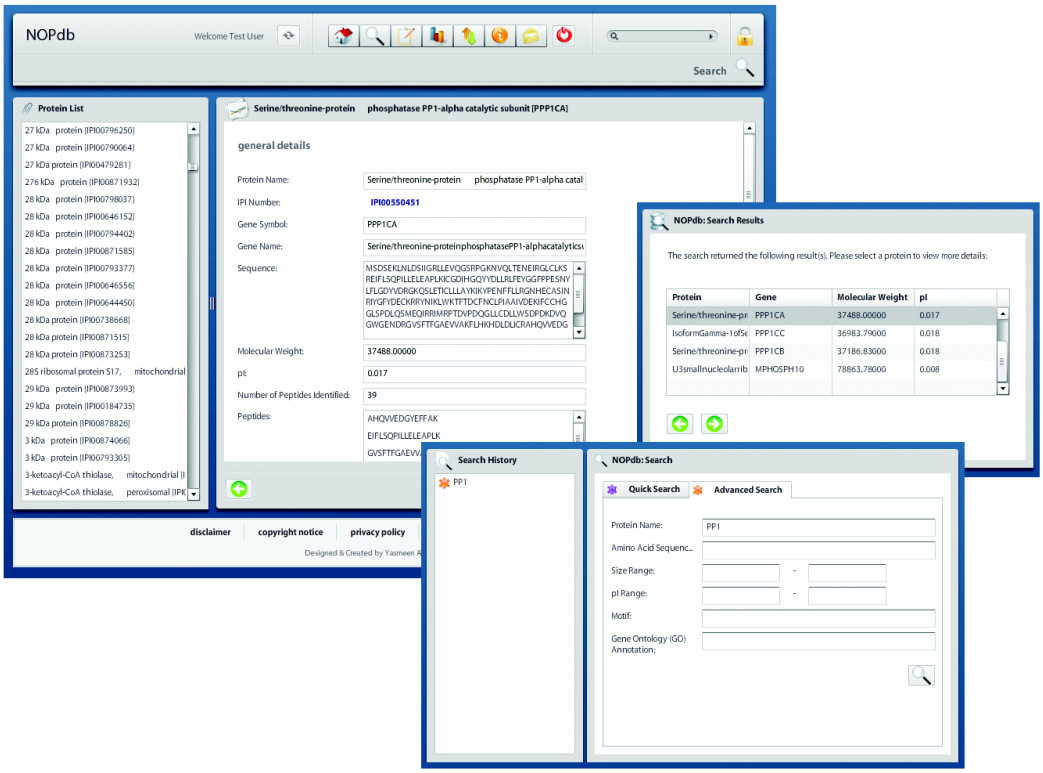


Figure 11: Snapshots of the NOPdb3.0 (http://www.lamondlab.com/NOPdb3.0/).

For illustration, the database was searched to identify a protein: Phosphatase 1 (PP1) isoform and here is shown an overview page for this protein documenting its sequence, peptides identified, etc.

The NOPdb3.0 provides information on multiple parameters, including protein name, accession number, gene symbol, gene name, sequence, molecular weight, isoelectric point (PI), peptides identified, experiments in which the protein was identified, motifs and gene ontology annotation.

3.4 Technical Implementation

The new NOPdb3.0 application consists of a multi-tier architecture, with the data storage, business logic and client interface as separate components.

3.4.1 NOPdb3.0 Databases

Version 3.0 of the NOPdb is an entirely new implementation using a fully relational design with major improvements over previous versions and additional functionality. The newly created database holds data of higher granularity, storing data at the peptide level as opposed to collated data on proteins. This higher granularity also means that results from new experiments can be directly uploaded to the database without prior processing, as the direct output from MS-based proteomics analyses is peptide data.

The data storage is implemented via a relational MySQL database. The database is structured (see Figure 12) to allow easy extendibility and maintenance in the future.

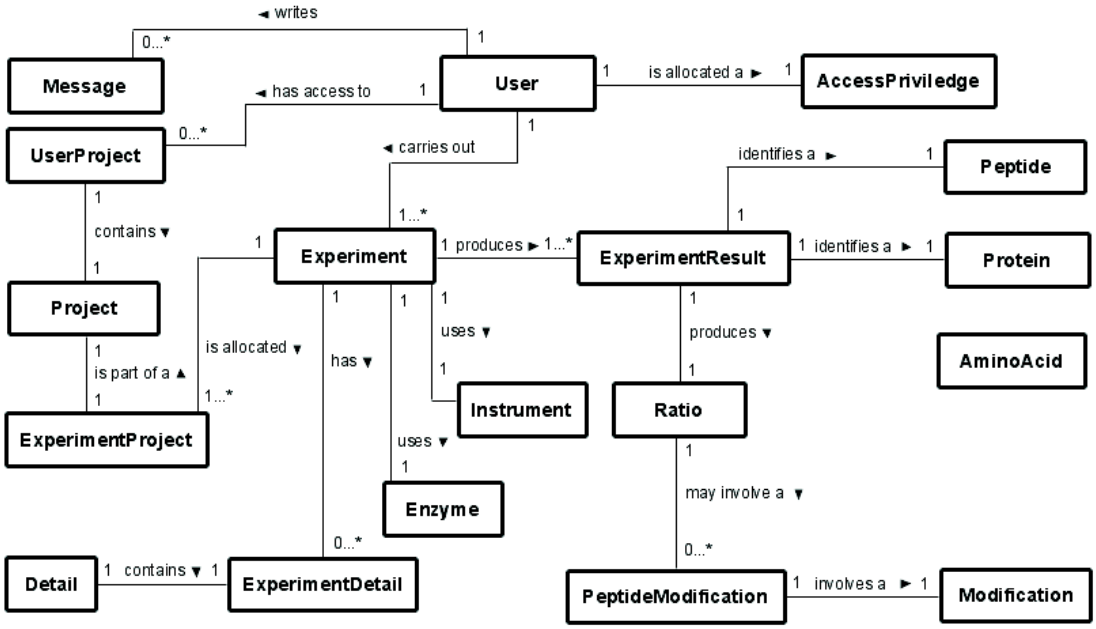


Figure 12: Entity-Relationship (ER) diagram for NOPdb3.0.

E-R diagram depicting the relationships between the tables present in the MySQL database implemented for the NOPdb3.0 Content Management System.

A number of database features were employed to ensure the security of data and to prevent SQL injection attacks. One of these features involved the use of Views that sit in a layer above tables. Access to the database was only permitted via these Views. Some of the View tables consist of aggregation of common queries, which increases the speed of querying, hence providing better performance to users. In conjunction with the View tables, specialised users were set up on the database. The database

users were granted restricted privileges to carry out specific operations on certain tables.

Normalisation of the database fields was carried out using the three normal forms defined by Edgar F. Codd (Codd et al., 1971), i.e. 1NF, 2NF and 3NF. This technique helped to eliminate anomalies from the database design and hence avoid any structural or logical problems. A definition for 1NF, 2NF and 3NF is provided below:

- **1NF**: A relation in which intersection of each row and column contains one and only one value.
- **2NF**: A relation that is in 1NF and every non-primary-key attribute is fully functionally dependent on the primary key (no partial dependency).
- **3NF**: A relation that is in 1NF and 2NF and in which no non-primary-key attribute is transitively dependent on the primary key.

The purpose of the business logic layer is to act as an interface between the client-side application and database. In order to extract useful data from the database, the business logic employs complex SQL queries.

A further two databases were also set up to store a local copy of the additional data from the International Protein Index (IPI) and Gene Ontology databases. The information required from these databases was found in file formats from the respective database websites. Scripts were then written to parse the data and store it in a relational format, which is linked to the NOPdb3.0 database. These data provide useful, further annotation to complement the data within the NOPdb3.0.

3.4.2 Application Programming Interface

All communication between the database and application has been implemented to pass through the custom made Application Programming Interface (API) to create data pages ‘on the fly’ using the custom API rather than serving static data pages, as in previous versions. The API acts as a security blanket around the database. All requests to carry out CRUD operations have to pass via the API, which ensures that the user has the appropriate privileges, via a unique API key supplied to each user. The API is implemented using the REST (Representational State Transfer) approach. REST is an "architectural style" that exploits the existing technology and protocols of the Web, including HTTP (Hypertext Transfer Protocol) and XML (Extensible Markup Language).

REST is simpler to use than the well-known SOAP (Simple Object Access Protocol) approach, which requires writing or using a provided server program (to serve data) and a client program (to request data). Within the NOPdb application, the REST technology is used to retrieve data and allow it to be read through a series of designated web pages that hold and describe the content in XML. Furthermore, the API to the NOPdb has the potential to allow the global science community to access the data held within the database, whilst still ensuring high security. It does so by providing a series of functions that cover a broad range of tasks that a user may like to carry out. These are planned out using the API design specification. Using these functions, external users are able to utilise the REST web service to create applications and/or websites that subscribe to data held within the NOPdb. Users require the URL (Uniform Resource Locator) for the page where the XML is located. They receive this information after they request access to the API and are then supplied with a unique API key and documentation. Users can then interpret the content data using the XML information and reformat it appropriately.

The API was coded to accept an API key from users and compare this to the database to determine which scripts a user can execute. Furthermore, the URL access to API scripts was implemented to use the mod_rewrite Apache server module. This module allows users to request data from user-friendly URLs, which are translated on the server side, using mod_rewrite, into a format that is more acceptable by the technical scripts.

3.4.3 Application Security

Increased security was a core focus of this development. The application itself is designed with three levels of access, to facilitate management and to prevent unauthorised use of the system. Users are provided with different levels of access according to their needs, which are seamlessly enforced by the application. This security ensures that the data remain accurate and the quality of the data is not compromised. Furthermore, this application creates a platform for the Lamond group to share their data with the wider cell biology community.

3.4.4 Client Side Interfaces

The application has the ability to interpret data and therefore aggregate it to provide metadata for proteins on a usable, graphical interface. The structure of the application

has been designed using the Model-View-Controller (MVC) design pattern, thus meaning that the functionality is separated from the overall look and feel of the application to ensure a more customisable solution.

The business logic and client interface can both reside on any Apache web server capable of serving PHP classes and the client interface, which is built in Adobe Flex. The Apache server had to be configured to allow scripts to run for a greater length than the standard of 30 seconds and to be able to handle larger file uploads. This was required to ensure the server could handle the processing of the large result files produced in the Lamond Laboratory. Furthermore, a cross-domain policy had to be added to the server to ensure that a flash player, running the client side application, could communicate with the PHP scripts residing on the server. Adobe Flex was chosen as it allows RIAs to be prototyped and developed rapidly, with the end product running across a wide range of client browsers.

Implementation began with the electronic prototype, used as a basis for further implementation. The mock functionality was periodically changed over to become fully functional with the API and database. During the implementation phase there were a number of features that required extensive thought, research and planning to implement in the best possible way – optimised for quick and efficient performance. Furthermore, a number of algorithms were developed to carry out common functionality in a well thought out sequence of steps. Some notable technical achievements include:

- An efficient searching algorithm to allow a user to carry out a generic search for specific proteins based upon a set of search criteria. This algorithm is used for the main search that is available in the application. The search algorithm created is intuitive as it is able to speed up the search, based on the search criteria provided. Some of the information provided in the search criteria spans multiple tables or even multiple databases. In order to ensure that minimal queries are carried out, the search algorithm uses different database views depending on which tables and databases need to be queried. Furthermore, if proteins have been eliminated from initial search queries, the algorithm will ignore these proteins when looking for matches with the remainder of the search criteria identified.

- The implementation of an API function that references a number of database tables to collate an amalgamated set of data based on one protein. It achieves this by referencing multiple views from different databases.
- The creation of a script that can efficiently process large result files (tens of thousands of lines) produced from lab instruments. These result files are provided in comma separated variable format (csv). The script is able to identify both errors and omissions in the data, which it then reports to the calling script in detail.

The application facilitates mining of stored data, with data being stored in a relational structure that is well documented. Thus tools can be built to search, analyse, read and understand the data. This mining capability is evident within the application interfaces, with the database being searchable by multiple parameters, including gene names, amino acid or nucleotide sequences, sequence motifs, or by limiting the range of isoelectric points and/or molecular weights. The database is also searchable by Interpro motif numbers (database of protein families, domains and functional sites) (Mulder et al., 2003, Bateman et al., 2004, Letunic et al., 2004) and by Gene Ontology terms (describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner) (Ashburner et al., 2000). Furthermore, the NOPdb3.0 application allows the researchers to visualise the data produced from experiments and enables cross analysis between experiments.

3.5 Discussion

Through investigating an existing application, used for archiving basic data on proteins, this study was able to create a unique, secure, extendable content management system, holding advanced nucleolar proteomics data. The newly created application includes a dynamic relational database, which is kept up to date by members of the Lamond Laboratory. The data are further annotated with information from reliable, recognised external sources on the web. The database and application have been implemented to communicate via a custom made API. This API acts as a security blanket around the database. In addition, an API provides external users with a portal to link into the Lamond Laboratory database and hence gain access to cutting edge research and results. This API provides the ability for users to create their own

websites and/or applications that represent the data being stored in the proteomics database.

The NOPdb3.0 allows secure access to the data, via three user levels. Users have control to enter new data into the system and carry out on-going management of the data. Furthermore, this application allows the Lamond Laboratory to share their data with the wider biology community, who can then benefit from access to the latest research results. The application also makes data mining a reality. The specialist techniques used to design and create the client interface enable it to dynamically generate layouts for content and allow easy access to data for Lamond Laboratory biologists. Through the interfaces, researchers can visualise the data produced from experiments, which allows for easier cross analysis between experiments to be achieved.

The database has been populated with different sets of experiments that identify proteins in purified preparations of human nucleoli. From the initial ~700 proteins identified in the previous iteration of the NOPdb, over 50,000 peptides have been identified contained in over 4,500 human proteins from purified nucleoli verified by multiple MS analyses in different cell lines, providing significantly enhanced coverage of the nucleolar proteome. The increased coverage of the human nucleolar proteome is illustrated by the fact that NOPdb3.0 now includes over 80% of ribosomal proteins, as opposed to the ~28% described in NOPdb version 2.0. It is estimated that NOPdb3.0 contains over 80% of the main human nucleolus proteins. The proteins in the database will be regularly updated as more experiments are performed in the Lamond Laboratory.

The continuous collaboration with the researchers from the Lamond Laboratory has resulted in the creation of a usable piece of software. The user-centred approach employed during this work involved closely working with researchers to continuously evaluate and obtain feedback on the software being created. Using this approach, user responses could be quickly incorporated into early prototypes resulting in fewer changes towards the end of the development. Feedback from the researchers suggested that previous experience of software development had not involved this user-centred approach, which they attributed to the success of this project. Furthermore, sessions with the researchers revealed that this project was simply

opening the doors and highlighting the possibility of many more developments in this field.

The nucleolar proteome database provides a basis for further work in creating a larger repository and interfaces that can handle all types of MS experiments, not only nucleolar datasets. Furthermore, the NOPdb3.0 highlights the potential of automating the storage of data in an electronic database and using it in conjunction with experimental metadata to carry out analysis. These concepts were explored to create new software described in Chapter 4: PepTracker - A Tool for Proteomics Data Management & Analysis.

Chapter 4: PepTracker - A Tool for Proteomics Data Management & Analysis

4.1 Summary

To date, biological science has relied heavily on the manual analysis of experimental datasets to interpret results. However, the combination of fast paced advances in both instrumentation and experimental techniques has resulted in the availability of complete genomes and the production of large datasets, for which manual analysis is no longer feasible. With the advent of high throughput methods for protein identification, huge volumes of data are being created via mass spectrometry. The volume and complexity of these data make it impossible to analyse them by manual inspection. These data are a major resource that requires new approaches to manage, analyse and store efficiently, for researchers working within the same institution and collaborating institutes. These problems are further enhanced by the complexity of the data, the non-consensus on data formats and absence of data standards. The issues mentioned here are outlined in detail within Chapter 1: Literature Review. Researchers are now heavily reliant on computer-based analysis and visualisation techniques for large datasets.

The Lamond Laboratory is one such group requiring new methodologies for the collection, storage, analysis, and visualisation of large datasets from proteomics experiments. The Principal Investigator of the Lamond Laboratory, Professor Angus Lamond, views the challenges surrounding proteomics data as an opportunity to innovate and drive the development of new software and techniques. With past success and access to state of the art equipment, the Lamond Laboratory makes an ideal test bed for developing new software that can be challenged by leading scientific protocols and researchers recruited with varying expertise, ranging from traditional biologists, chemists, mass spectrometrists to pathologists.

Through the development of new software the Lamond Laboratory aims to discover patterns and trends, which were not previously visible. This chapter describes an effort to tackle these issues, with the goal of making advances in the overall management of the datasets produced by experimentalists, through the development of web-based visualisation tools that have access to a rich amount of data from past years of

experimentation. These tools are aimed at providing insightful analysis through interactions with individual datasets, as well as allowing for comparisons of data produced by different researchers, using both similar and different experimental methods and thereby helping to promote new collaborations and cross-fertilisation of projects.

This project has resulted in the creation of software, PepTracker (http://peptracker.com), which allows robust data management and analysis capable of dealing with high throughput quantitative data from MS experiments and the corresponding metadata. PepTracker provides a suite of visualisation and analysis tools to facilitate complex data mining tasks, including the objective normalisation and comparison of data from separate experiments.

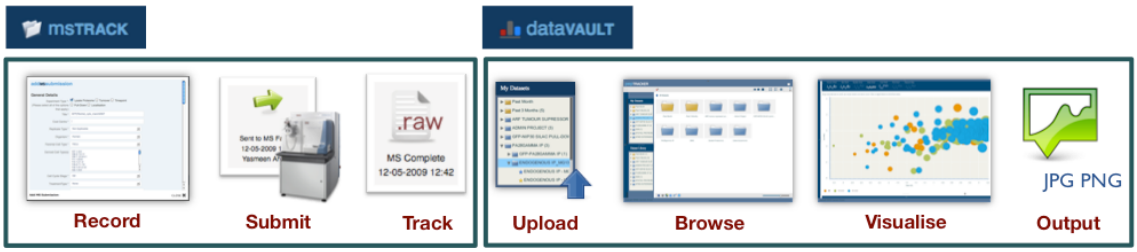


Figure 13: PepTracker workflow.

The PepTracker software consists of three main components: MsTrack - an integrated Laboratory Information Management System (LIMS), DataVault – sophisticated data storage, visualisation and exploration for quantitative proteomics data and ProteinLibrary – a protein search and management view that allows for querying of all datasets and creation of specialised protein groups. Chapter 4 describes the PepTracker software, focusing first on its three main components (sections 4.2-4.4), its implementation (sections 4.5 and 4.6) and finally on how data security and quality control are ensured (section 4.7).

4.2 MsTrack – Laboratory Information Management System

The PepTracker system has been implemented to incorporate a Laboratory Information Management System (LIMS). This LIMS provides recording and tracking of samples, automated sample submission and intelligent data management. This functionality ensures comprehensive linkage between the various metadata, raw files and processed data (see Figure 14). Thus, researchers can retrace their steps and re-analyse samples from any stage in the processing pipeline.

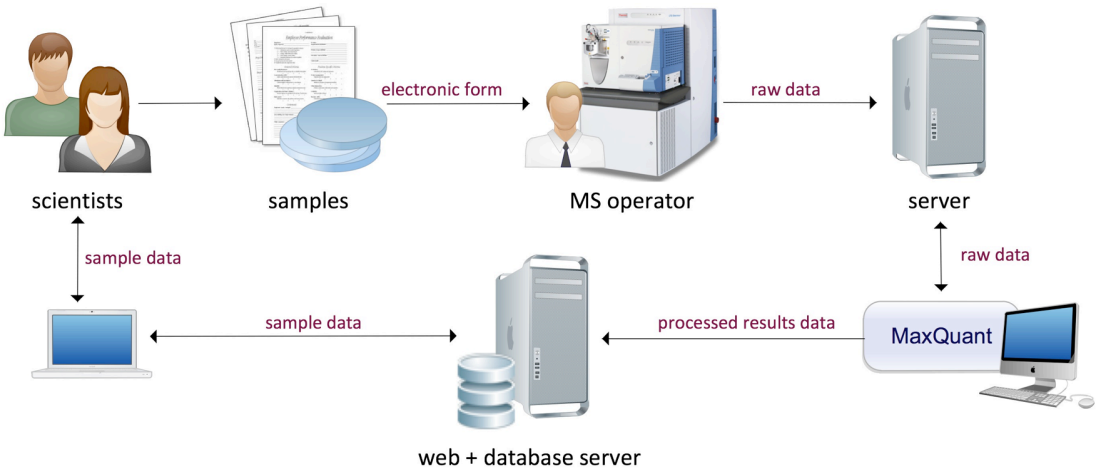


Figure 14: MsTrack workflow.

The LIMS part of the system records extensive metadata associated with the experimental procedure used to create samples, as well as data regarding the mass spectrometry process itself. The metadata collection was meticulously planned to ensure that it was comprehensive enough to allow in depth analysis of data at a later stage. In order to ensure data analysis can be carried out effectively it is imperative to know the samples from which the data were generated. For example, when carrying out data mining that involves collating historical data to analyse trends in immunoprecipitation experiments (see Chapter 5: Multidimensional Analysis with IP Experiments), it is important to be able to identify a suitable subset of mass spectrometry data that represents immunoprecipitation experiments, which can only be done effectively through filtering well formed metadata. Hence, in addition to simply collecting metadata, particular time was focused on ensuring it is collected in a structured manner, for example using drop down lists where possible.

The metadata collection is carried out via a wizard interface that sections the data to be collected into logical steps that are more comprehensive for a researcher wishing to enter data. The steps in the wizard are customised dynamically based on the data entered by a researcher. For example, if a user specifies that they have carried out a SILAC experiment, an additional SILAC step is added into the wizard (see Figure 15). Additional methods, such as the requirement to enter data using drop down lists, ensure researchers are consistent with data entry, avoiding mistakes such as spelling errors, grammatical differences, variances in expression and enforcing the use of a standard nomenclature.

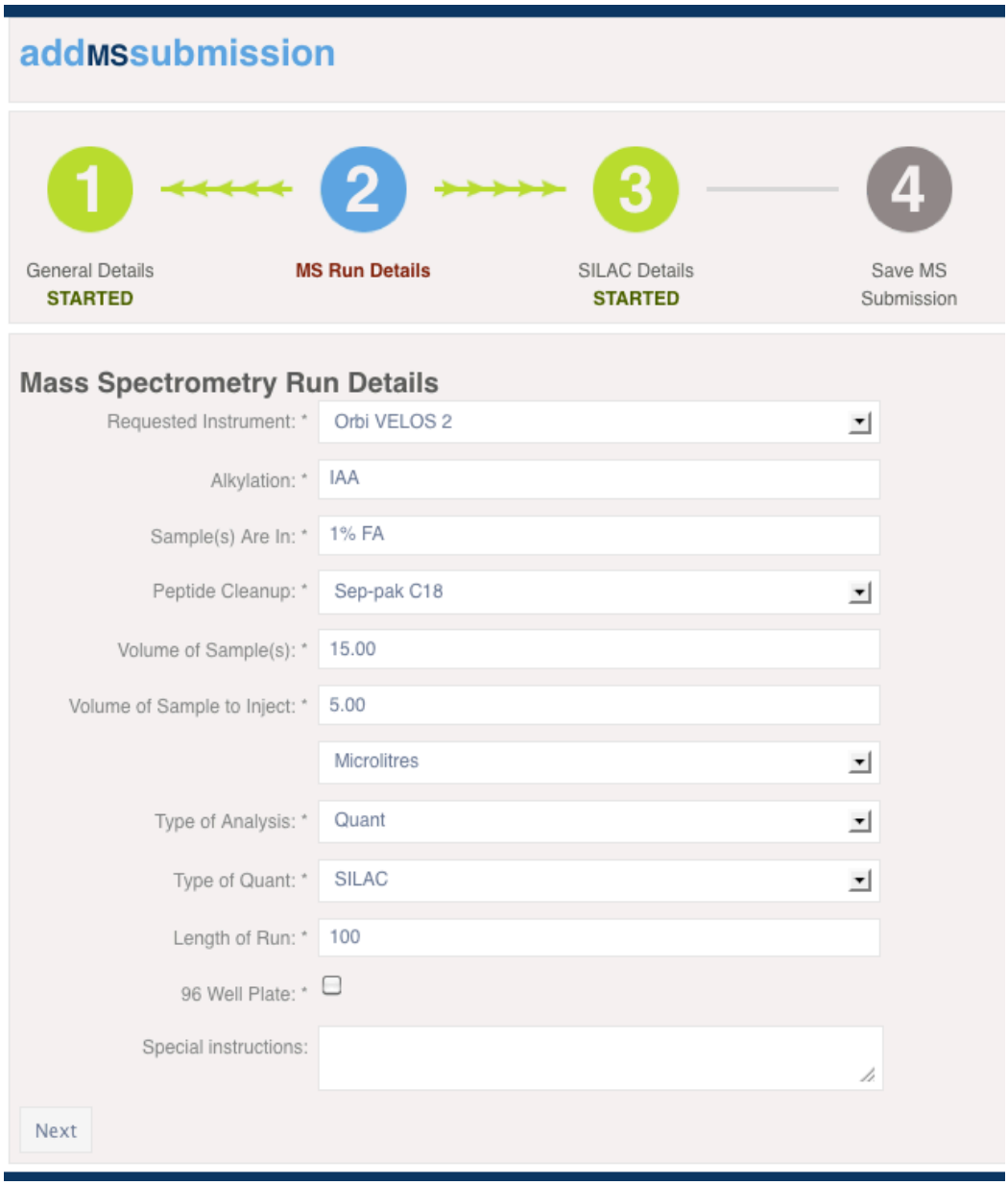


Figure 15: MS submission wizard interface.

Furthermore, to ensure consistency of data entry, the LIMS system incorporates a reagents database. This part of the LIMS system manages data on antibodies, plasmids, cell lines, siRNAs and chemicals within the laboratory. When creating a new experiment, researchers can choose items from the reagents database rather than entering free text. Furthermore, metadata entry was customised so that PepTracker can be comprehensive with regards to the information stored about experiments. Hence, the data requested varies dependent upon the type of experiment carried out. For example immunoprecipitation experiments require information to be entered about buffers, beads and the protein pulled down. Also metadata collection may be customised if a protocol is specialised, such as SILAC, which requires input of

information about labelling used and description of each label condition. A full specification of the metadata collected is available in Appendix B.

Users, wishing to add, update or remove entries, can amend all lookup lists in PepTracker. However, the developer reviews any changes to these lookup tables on a monthly basis. This strategy was chosen as it was realised that proteomics experiments are constantly evolving and improving, which means researchers will need new options in the lookup tables. By allowing user additions dynamically, the software encourages users to be thorough in their data entry rather than entering incorrect data or omitting data because their ideal selection item is not in the list.

The metadata collected regarding sample preparation is used by PepTracker to create an electronic sample submission form for the Fingerprint Proteomics Facility at the University of Dundee (http://proteomics.lifesci.dundee.ac.uk/). The electronic submission is sent to the MS facility upon researcher request, leaving the researcher to submit their physical samples, which must be labelled as per PepTracker convention. In consultation with the in-house Fingerprint Proteomics Facility, the workflow is implemented to ensure all details of MS experiments conducted in the laboratory must be entered into PepTracker before MS analysis is possible. Importantly, this ensures that every MS experiment performed is always contained in the PepTracker database. PepTracker can also manage and integrate “legacy data”, entered either from older experiments from the laboratory, or from other researchers who do not have access to the LIMS component.

After MS submission, the MsTrack component monitors the status of the submission. It does so by initially logging the researcher, date and time of an electronic MS facility submission and using this information to keep the researcher updated with regards to the progress of the submission. To determine when samples are being run, PepTracker continuously polls the proteomics server to identify the presence of raw files associated to the samples. It identifies relevant files using the unique name allocated to each sample submitted to the MS facility. When PepTracker is able to identify the presence of the first of these files, it logs the date and time, indicating the starting point of the MS analysis of the samples. Once all submitted samples have been linked with a raw data file on the proteomics server, PepTracker once again updates the status of the submission by logging the data and time when the MS submission was

identified as being complete. At this stage PepTracker sends an automated email informing the researcher that the raw data files are available for further processing. Depending on where the raw files are located on the proteomics server, PepTracker is able to determine which instrument was used to run the samples. Moving forward, PepTracker maintains a link, providing online access to the resulting raw files in the future for re-analysis if this were deemed necessary, for example if updated software becomes available.

Once the raw data files are available, researchers can download these files from PepTracker through a one-step click. After carrying out the standard quantitative analysis using 3rd party software: Mascot and MaxQuant, researchers can upload their MaxQuant data to PepTracker for further data analysis and visualisation. This includes user-selected options to automate downstream analysis and visualisation procedures, dependent upon the type of experiment being performed. For example, pull-down analyses of protein-protein interactions can be selected for automated data plots and identification of contaminants and normalisation of data points with reference to the Protein Frequency Library (see Chapter 5: Multidimensional Analysis with IP Experiments).

Currently, the LIMS part of the system does not automate the data processing stage. Due to the evolution in software this is a task that could not be easily automated. However, as the software matures and becomes more stable this is a possibility for the future development of PepTracker. Although not essential, this will streamline the workflow and result in a fully automated system for the researcher, going from submitting samples to the MS facility through to receiving output files from MaxQuant, with quantified peptide identifications entered and managed within PepTracker.

4.2.1 Tag Cloud

A tag cloud was implemented in PepTracker, using the keywords that are specified for MS submissions by researchers. The size of the keywords on the tag cloud is determined by the frequency of their use in MS submissions. This provides a quick visual as to the types of experiments being carried out in the laboratory and any ‘hot’ topics (see Figure 16). Current users of PepTracker thus often employ gel filtration and fractionation in the preparation of their samples, work with worm and U2OS cells,

most often run their samples on the VELOS mass spectrometer and within the Lamond laboratory there is a particular interest in the proteins SMN and MRFAP1.

When entering keywords, users can choose to either select three words from an existing list, populated from all keywords entered into PepTracker, or users can enter new words. This type of keyword selection shows problems as users often simply select three keywords from the existing list for ease, rather than entering new words. It is realised that this will develop over time as an improved vocabulary of keywords is built up. In addition, PepTracker has been updated to allow researchers the option of selecting as many keywords as they want rather than restricting to only three words.

Figure 16: Tag cloud generated from MS submission keywords.

4.3 DataVault – Storage, Visualisation & Exploration of Quantitative Proteomics Data

The DataVault provides researchers with the means to upload quantitated data, which is parsed and stored in a data warehouse. These data are then presented to researchers via an interactive interface that allows exploration and discovery tasks to

be carried out. The DataVault has been created with a desktop appearance and interaction. The focus of the interface is to maximise interactivity with data and provide easy access to features that aid data exploration.

The overall DataVault interface is split into various panes that separate functionality into logical sections. The two main views on the DataVault are the browser view and data view. The browser view provides the ability to upload and store MaxQuant datasets and implements an organised view, browsing functionality and selection of multiple datasets. The data view focuses on the visualisation and exploration of one or more datasets. These views are further enhanced by menu bars that provide additional features such as Full Screen mode to allow maximisation of screen usage for the exploration of data (see Figure 21).

4.3.1 Data Storage

Once raw data have been processed using MaxQuant to generate quantitative data, researchers can choose to upload their resultant dataset to PepTracker for downstream analysis, visualisation and exploration.

Considerable effort has been devoted to optimise uploads of MS data so that PepTracker can handle the varying sizes of datasets in an efficient manner. Datasets can vary from a few hundred megabytes to a few gigabytes. PepTracker handles the upload of these datasets by providing a form for researchers to specify files from their MaxQuant output plus a description of the dataset. The output files from MaxQuant used in the upload are listed in Table 4.

File Name	Description
parameters.txt	Logs the parameters used in the MaxQuant search.
experimentalDesign.txt	Documents the fractions, experiments and raw files that are represented in the dataset.
evidence.txt	Combines the information obtained about the identified peptides in a mass spectrometer run, logging a row for each occurrence of a peptide.
modificationSpecificPeptides.txt	Contains aggregated information on the identified peptides in the processed raw files, logging only one row for each modified version of a peptide.
peptides.txt	Contains aggregated information on the identified peptides in the processed raw files, logging only one row for each peptide.
proteinGroups.txt	Contains information on the identified proteins in the processed raw files. Each single row contains the group of proteins that could be reconstructed from a set of peptides.

Table 4: MaxQuant Output Files.

The experimental design file is important as it is used to link back to the original metadata used in the submission of the MS experiment. PepTracker creates this link through matching raw file names from experimental design file to sample names that were originally logged in the PepTracker database at the time of MS Submission. The remainder of the files, required in the MaxQuant data upload, contain the major quantitative information that was generated by the mass spectrometer and used by researchers to generate hypotheses and draw conclusions from data.

During the dataset upload, PepTracker uploads the MaxQuant data files to a temporary folder on the PepTracker server and then processes the files as a background task using the PepTracker scheduler (see 4.6.4 PepTracker Task Scheduler, Figure 17). This scheduler has been created to manage tasks on the PepTracker server in the background so that researchers on the front-end are not restricted to keeping a web page open for an extended period while the data upload and processing occurs. Instead researchers can navigate away from the upload dataset page and the scheduler automatically notifies them by email once their task is completed. This scheduler also aids in distributing the processing required if multiple researchers attempt to upload datasets at the same time.

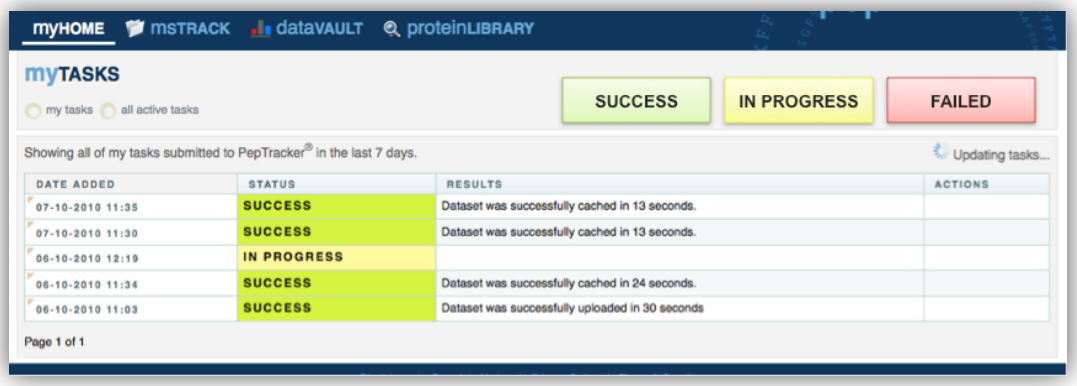


Figure 17: PepTracker scheduler for handling data uploads.

4.3.2 Browser View for Dataset Management

One of the important needs highlighted by researchers was the easy navigation and management of datasets generated from MaxQuant. Hence, PepTracker has been implemented to provide a variety of views to browse the datasets available in the PepTracker database. These views include a list view or customised icon view (see Figure *18*). The icon view allows for quick recognition of different types of experiments and datasets.

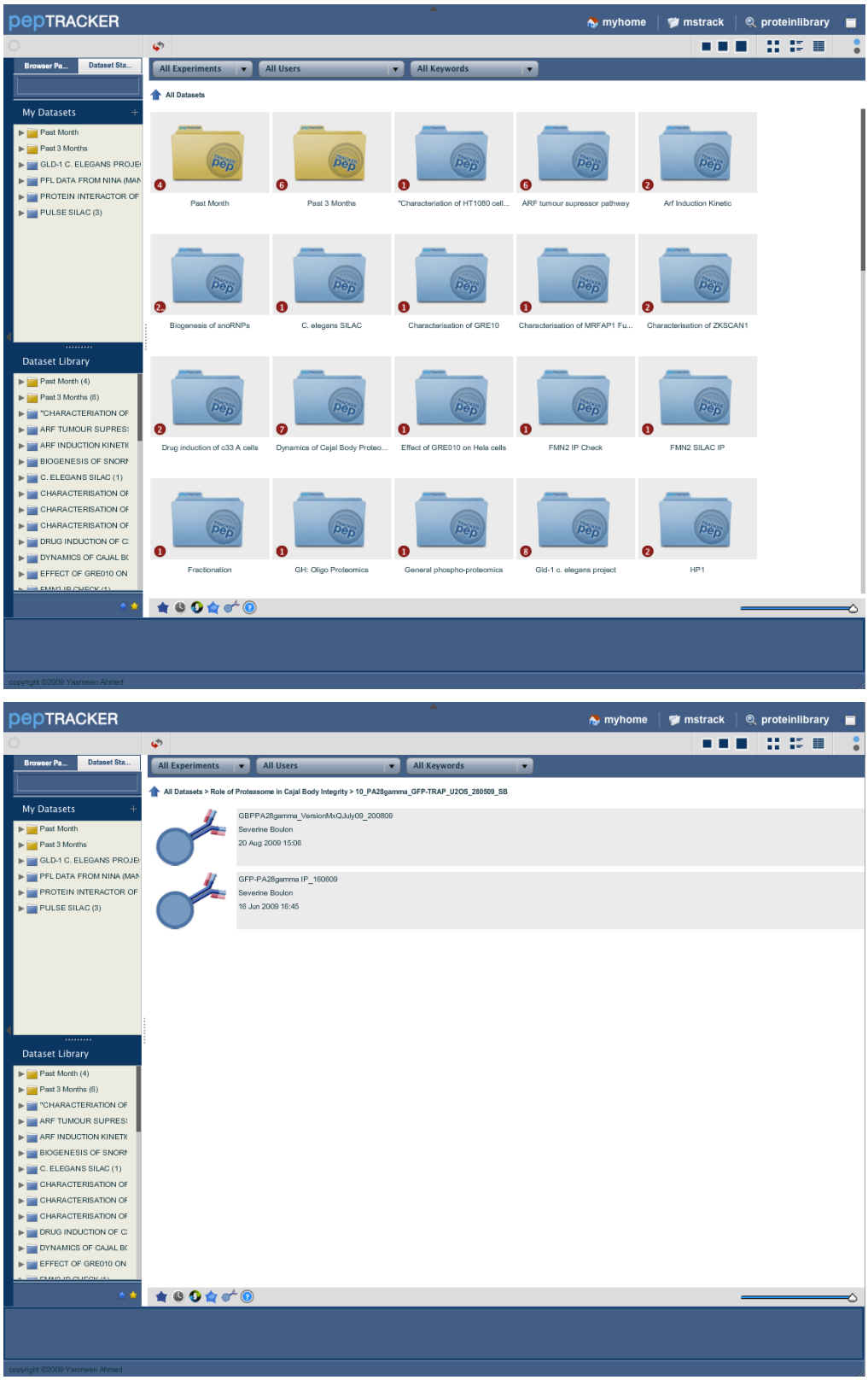


Figure 18: DataVault browser views.

The browser view also provides smart folders that allow easy access to the most recently uploaded datasets. Furthermore, researchers can set up their own smart folders containing a set of specified datasets of their choosing. A search feature is also available, which filters datasets dynamically based on keywords entered into a search

box. In addition, researchers can filter the datasets viewable via drop down lists that allow specification of researchers, keywords and experiment name etc.

4.3.3 Data View for Data Visualisation and Exploration

In order to provide additional benefit of having datasets stored in PepTracker, researchers felt it would be invaluable to have methods of viewing their data in novel ways. This involved providing advanced interactivity with datasets, functionality that mostly is currently unavailable with other proteomics software.

A set of tools has been created within PepTracker for the convenient visualisation, statistical analysis and comparison of datasets. Advanced tools are currently available for second-generation proteomics experiments, including protein pull-down analyses (see Chapter 5: Multidimensional Analysis with IP Experiments), spatial proteomics and pulse labelling/turnover studies (see Chapter 6: Spatial Localisation & Turnover Analyses).

Interface Evolution

Initially the PepTracker visualisation functionality was implemented using Google's Visualisation API (see Figure 19). This API provides a series of charts and gadgets that can be incorporated into web pages. The charts used HTML5/SVG technology to provide cross-platform interactive visualisation of data. The data source was supplied to the charts using JavaScript code.

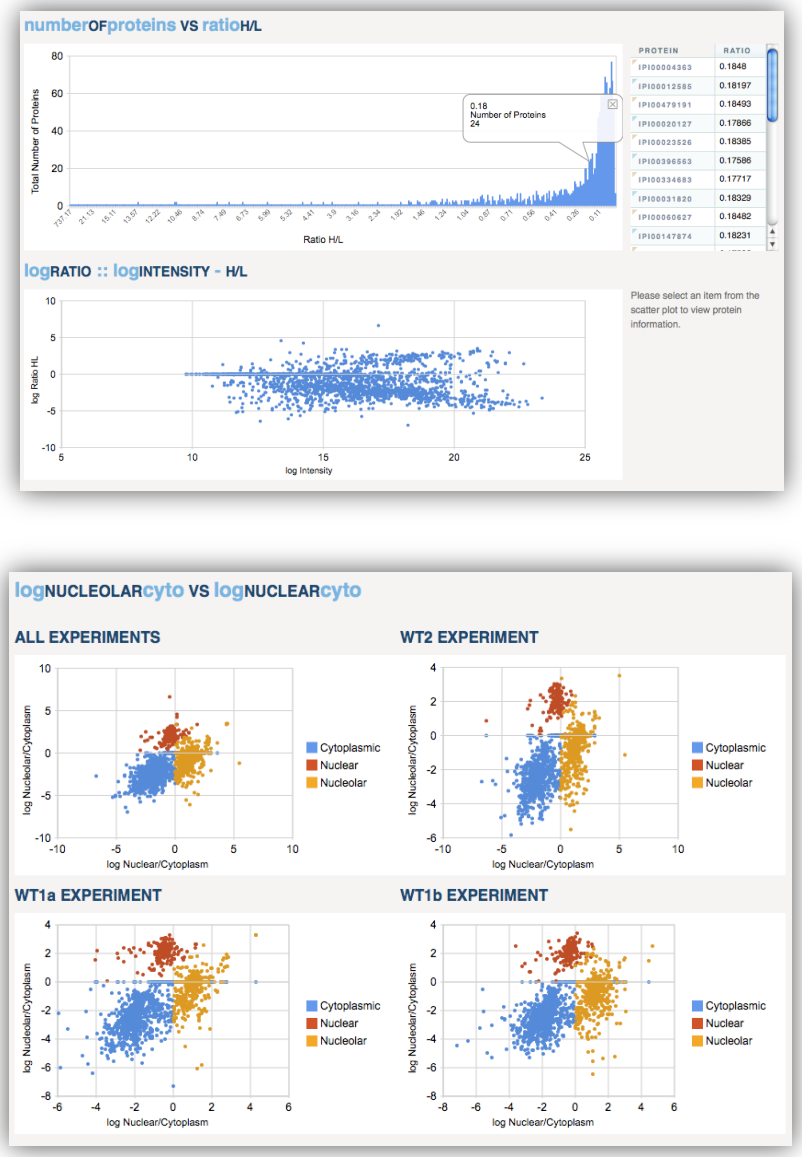


Figure 19: Early PepTracker interfaces created using Google Visualisation API.

This solution was effective as it provided researchers with the ability to navigate and explore their data using simple mouse manoeuvres. However, drawbacks to this technology included the considerable development time required to integrate the charts into a wider solution and the limitations in customisation of the charts. Researchers also highlighted the restricted interactivity with the data. Furthermore, researchers wanted a solution that provided flexibility with the data fields used on the visualisations and advanced features such as slider-based filtering of chart data. Hence, new solutions were explored and the final outcome was the use of Adobe Flex technology to create an integrated web and desktop based viewer that provided a more immersive experience for the researchers.

Visualisation Interface

The data view on the DataVault provides a variety of chart and graph types to explore data (see Figure 20). These include:

- Bar Chart
- Column Chart
- Bubble Chart
- Area Chart
- Line Chart
- Plot Chart
- Network Map

The charts and graphs have the ability to detect and plot various experiment groups within a dataset as separate series on a chart or graph of choice. Furthermore, researchers can select multiple datasets and plot one dataset variable against a second dataset variable on the same graph or chart.

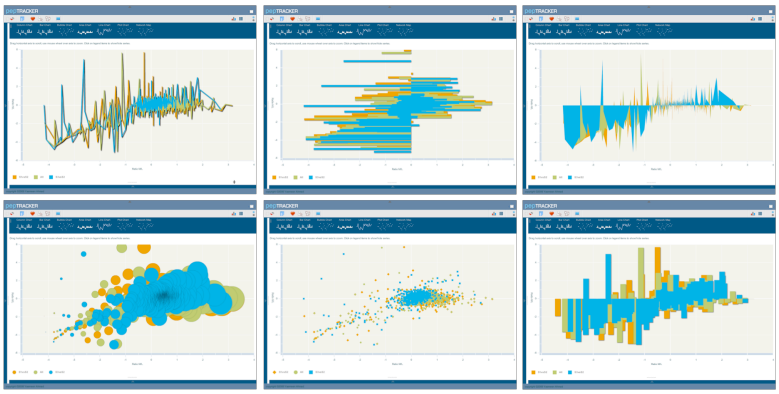


Figure 20: DataVault graphs and charts.

PepTracker provides a range of interactive features with the chart visualisations including:

- Customisable x, y and z-axis to plot user selected variables.
- Scrollable and zoom-able axes.
- Option to show/hide series.
- Mouse and slider enabled range selection that can be used to zoom into a chart or to create a custom protein group.

- On click context menus that provide various options including easy access to protein information pages.
- Customisable colours for single data points or series.
- Drag and drop, dynamic overlay of protein groups on a dataset.
- Mouse hover tooltips that can be used to identify proteins.
- Filter options, including contaminants, reverse proteins, PFL frequencies etc.
- Customisable chart colours, fonts and line thicknesses.
- Output to JPG and PNG formats.

These interactive features allow intuitive navigation of datasets and provide the ability to understand data more quickly. Through interacting with the data at the protein level, researchers can drill-down to discover interesting clusters or individual proteins for further exploration. From the researcher experience of interacting with the data (see Figure 21), it is intended that a better feel for the data can be obtained and hence, time and effort can be focused on hypotheses that are grounded in experimental evidence.

Figure 21: Interactive DataVault interface.

The visualisation interface also provides the ability to create dynamic network maps (see Figure 22).

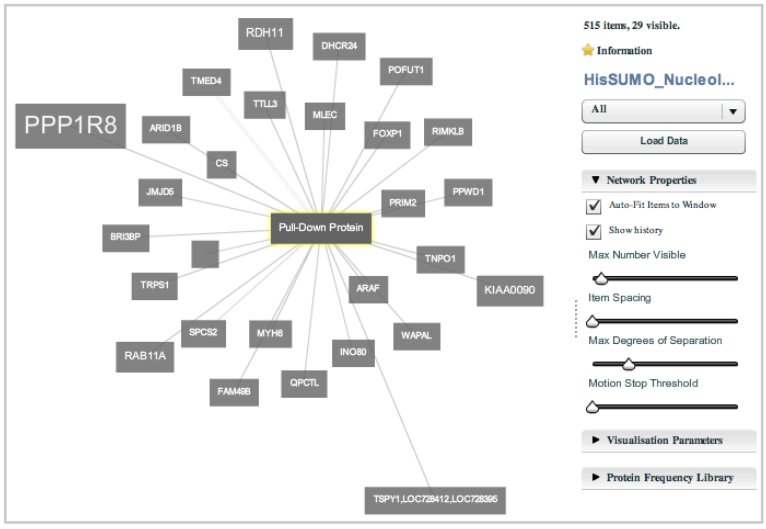


Figure 22: PepTracker protein network map.

The aim of these maps is to allow researchers to visually detect which proteins might be of interest to them. In order to make this task easier and quicker the network map is customisable so users can control various parameters such as:

- Line Thickness/Alpha/Colour
- Node Alpha/Colour
- Text Size
- Distance from Root Node
- Node Spacing/Visibility
- Number of Nodes Visible

Each of the parameters mentioned above can be allocated a variable from the dataset, such as ratio, intensity or peptide count, in order that they each visually represent an aspect of the dataset (see Figure 23).

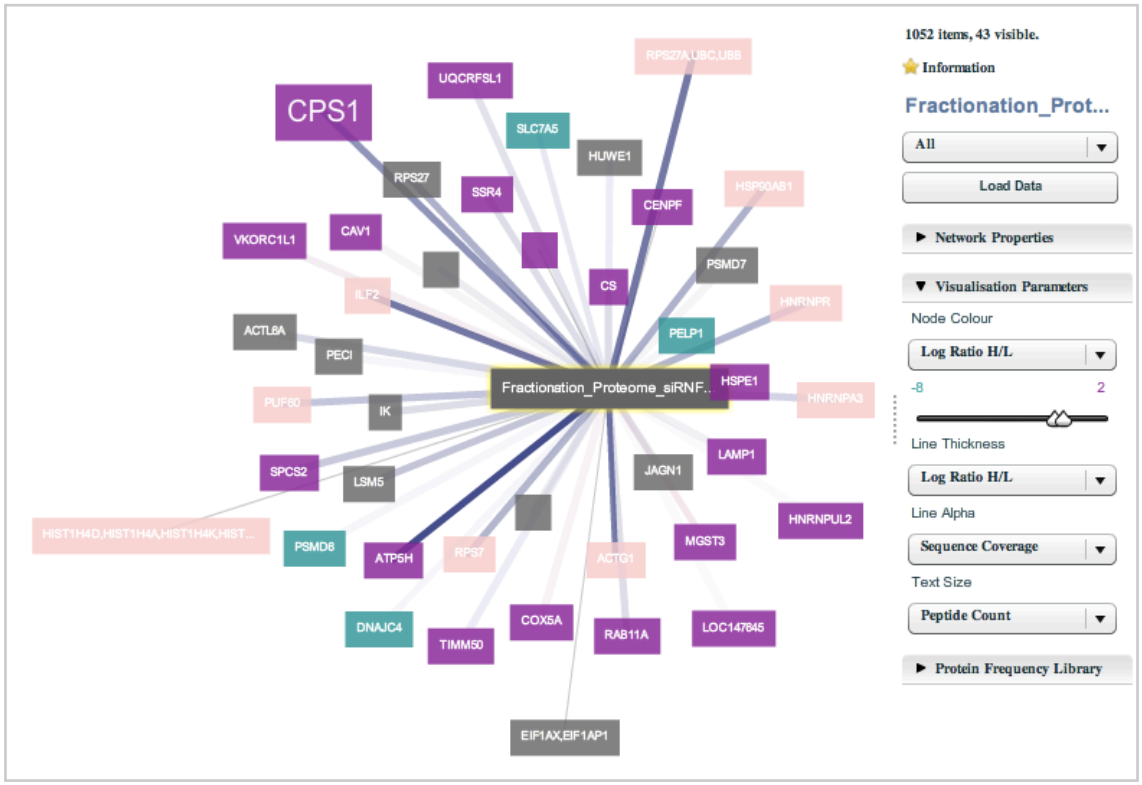


Figure 23: PepTracker customised protein network map.

4.4 ProteinLibrary – Protein Search and Specialised Protein Group Management

The ProteinLibrary section within PepTracker provides a simple search interface to allow researchers to look for a specific protein in the PepTracker database. It also provides functionality to create specialised protein groups.

4.4.1 Protein Search

The protein search within the ProteinLibrary provides researchers with the ability to search for a protein using UniProtKB identifier, gene name or protein description. Once searched, the ProteinLibrary interface is able to provide a quick overview of the datasets that were found to contain the protein of interest. From these results, links are provided to access further information regarding the protein identification within a specific dataset.

One of the visualisations provided by PepTracker is the Protein Peptide Alignment Map (see Figure 24). This diagram is able to quickly convey the reliability of identification by indicating the sequence coverage of peptide identifications. The map consists of two parts, a linear block representation (grey bar representing a protein with superimposed shades of red blocks representing peptides in the correct position) and a

sequence representation (amino acid sequence of the protein, with peptide sequences highlighted in shades of red). On both representations of a protein, all peptides within the sequence that have been identified and quantified by MS are highlighted in colour, which quickly conveys the peptide distribution and sequence coverage, and colour coded to illustrate conveniently the variation in SILAC ratios for each peptide quantified.

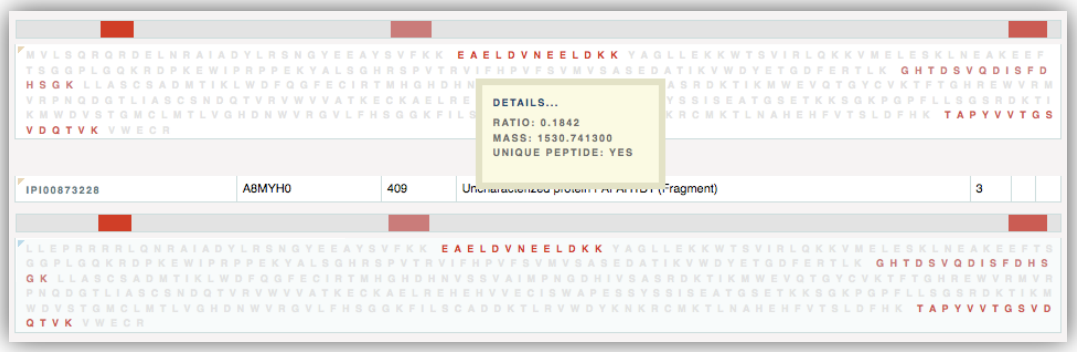


Figure 24: Protein peptide alignment map.

Finally, some reporting views have also been implemented. These are created in the form of information sheets. Each information sheet relates to a record that was stored in the MaxQuant results data output. These information sheets are linked so that researchers can navigate between data from the four separate MaxQuant entities, i.e. evidence, modified peptide, peptide and protein group. The protein group information sheet provides the aggregated form of data from the other three entities (see Figure 25).

Figure 25: Protein group information sheet.

4.4.2 Protein Group Definition and Enrichment Analysis

Another feature of the ProteinLibrary is the ability to manage protein groupings. During the analysis of mass spectrometry data, one approach employed involves grouping proteins based on a variety of parameters, such as their function, behaviour etc. This strategy allows researchers to deal with the complexity of the data and provide guidance as to which proteins should be further investigated. By keeping track of groups of proteins that are of interest, researchers can monitor their behavioural properties between experiments.

In order to aid in the analysis of datasets, PepTracker provides functionality to setup protein groups that can be overlapped with datasets. Currently there is the option to create three different types of protein groups:

- Sub-selection of a dataset.
- Researcher-defined group (defined by UniProtKB Identifier).
- Globally defined group (defined by UniProtKB Identifier or Gene Ontology).

These groups can be used in conjunction with the visualisations to show the intersection of the group with the current dataset. Furthermore, especially in the case of gene ontology defined groups, researchers can view the enrichment of a particular group in each dataset, for example the enrichment of nuclear proteins (defined by the 'nuclear' gene ontology term). This enrichment is calculated based on intensity values produced by MaxQuant, which provides a more accurate enrichment value than the traditional approach of simply counting the number of proteins found in the dataset that overlap a particular set of known proteins.

4.5 PepTracker Desktop Client

The DataVault component of PepTracker has been further extended to run as a desktop application with web connectivity. The Adobe Flex code base that makes up the DataVault is used and extended within Adobe AIR (Adobe Integrated Runtime), which provides a cross-platform runtime environment for building rich interactive applications that can be deployed as desktop applications.

As a desktop client, the DataVault can be downloaded and installed by a researcher on their machine and then used for visualisation and exploration of datasets. This method, compared to a web browser, is more robust and powerful as the application can make use of the full resources available on a user's machine as opposed to the restricted access provided by a web browser. The desktop client can be further enhanced in the future to implement a local database that can provide offline access to datasets.

4.6 Technical Implementation

4.6.1 System Architecture

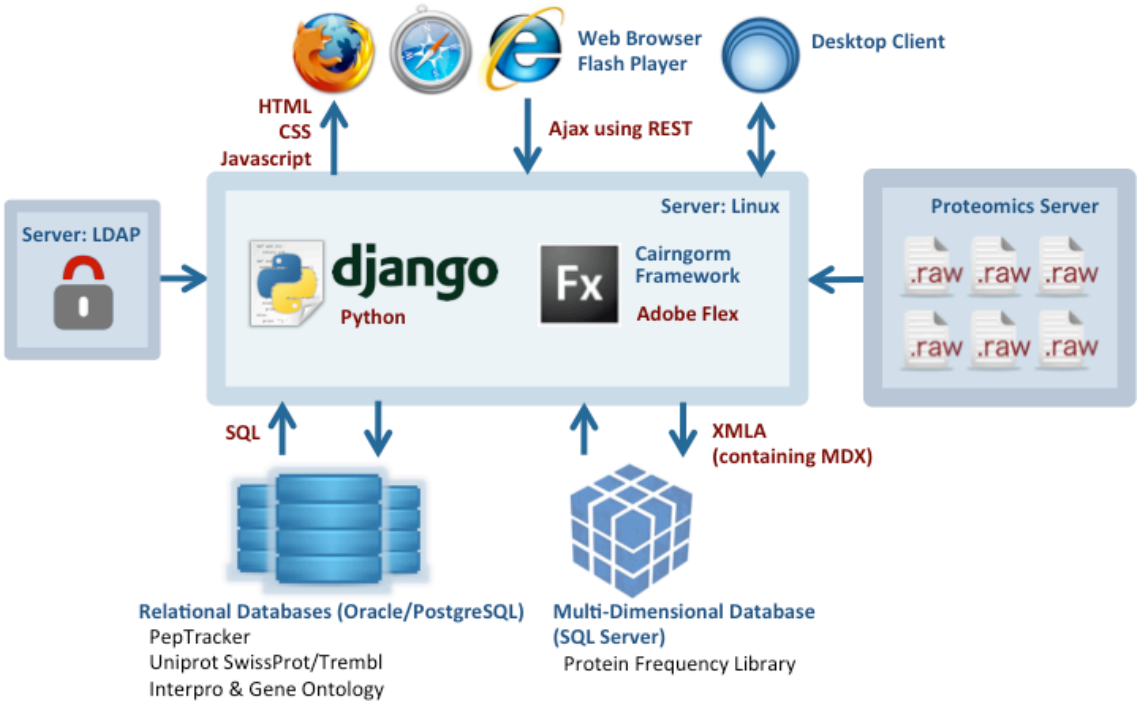


Figure 26: PepTracker system architecture overview.

4.6.2 Database Development

PepTracker contains two different types of databases, relational and multi-dimensional (described in 5.4.2 Multidimensional Database).

The relational database was originally set up using a MySQL database engine (version 5.0.77). MySQL is a multithreaded, multi-user SQL database management system that is fast and robust with a good feature set. The administration and security setup of a MySQL database are effective and not over complicated to implement. Even though MySQL is a very popular, open source web database, reliable and easy to use, it cannot match up to the performance and the variety of advanced tools and functions that are available in a larger database engine, such as Oracle. Oracle is much more versatile and can run and handle more transactions compared with MySQL. As a project grows, both in terms of functionality and/or users, Oracle provides additional features such as SQL transactions, stored procedures and data transformation services, which become important to ensure good performance of an end application.

During the development of PepTracker, the functionality and demands from users resulted in the decision to move from the original MySQL database engine to an Oracle

instance (Oracle Database 10g Enterprise Edition Release 10.2.0.5.0). This Oracle database engine is able to handle the data capacity growth and ensure performance for end users. However, due to the expense associated with an Oracle license, it is understood that not all institutes will readily have access to an Oracle license and so it was decided that an open source, alternative backend database option should be maintained. The database engine chosen for this task is PostgreSQL as it is easy to maintain, yet provides some of the advanced functionality available in Oracle.

The relational database holds all datasets uploaded to PepTracker and also stores extensive metadata describing the conditions under which experiments were carried out and parameters associated with the experiments, such as cell type, organism, extract, type of beads, machine, date, user, antibody and treatment etc. Furthermore, local copies of the frequently accessed components of the publicly available UniProtKB database (Apweiler et al., 2010) are included in the relational PepTracker database. These UniProtKB tables are periodically updated (every 4 weeks) to coincide with the UniProtKB releases.

The main PepTracker application communicates with the relational databases using SQL. However, various databases will use their own SQL dialect, hence tying the application to a specific database engine. Projects, such as SQL Alchemy, attempt to remove this dependency by providing a SQL toolkit and object relational mapper that is non-database specific for high performance database access. This provides an abstracted layer above the database, that the application can use rather than directly interfacing with the database. This software toolkit was adopted in PepTracker to make database selection and communication more maintainable and extendible. The major benefit of using SQLAlchemy to communicate with the relational PepTracker database is that the backend database engine can be changed without having to alter the server side code that communicates with the database. This was particularly useful when the decision was made to move the PepTracker database from the MySQL engine to an Oracle database.

The relational database is normalised to third normal form, as defined by Codd in 1971 (Codd et al., 1971), to eliminate data redundancy and inconsistencies within and between tables and to make the database more maintainable. This normalisation results in relational databases modelling reality by breaking the data into one or more

sets, each of which represent a class of real-world entity, for example protein, experiment, user etc. This method creates a data store that is optimal for storage and transaction based operations. However, more complex relationships and larger volumes of data can lead to slower analytical performance, especially as the analysis becomes more complex, hence additional tables are implemented that contain an aggregated copy of denormalised data to act as a cache for fast access.

4.6.3 Server-Side Setup

The main PepTracker server is a HTTP (HyperText Transfer Protocol) server that is implemented on a Virtual Machine running Linux (CentOS Release 5.3). This HTTP server is able to communicate with web browser clients by understanding HTTP requests that are sent to it and forming valid HTTP responses that the client machine can display as HTML (HyperText Markup Language) pages on a web browser. Hence, the server has to store all code that is served to researchers when requests are made through a browser.

Any software that is put into production needs to be fixed, maintained and updated. Each of these aspects requires either the modification or addition of source code. It is therefore of the utmost importance that key software systems (such as a data management and analysis systems) are written in clear, understandable and documented code. The choice of programming language was Python as it was understood that all code would need to be carried through for many years, potentially having to deal with changes in programmer staff over that time period. Python benefits from being a very robust, object-oriented language with highly desirable cross-platform capabilities. This means that the compiled code itself can be distributed and run on diverse platforms and architectures without any problems and with only minor (if any) differences. Finally, the popularity of Python and the elegance of its object-oriented design have made it a favourite for programmers, especially those in the bioinformatics field.

The majority of the PepTracker server-side code base is written in Python (version 2.6.1) using the Django framework (version 1.1.1). Django provides a high-level Python Web framework for rapid development and structured, clean, pragmatic design of code. The Django code is structured using the Model-View-Controller (MVC) design pattern so that the database, business and client logic are separated. Django has an in-

built object relational mapper that allows definition of data models and provides a dynamic database access API. Django's template system provides a powerful template language that is extensible to separate the application interface design from the Python code.

As well as making use of the in-built Django database access API, raw SQL statements had to be constructed for more complex queries. In order to ensure the code remained extensible and not tied down to one database engine, SQLAlchemy (http://www.sqlalchemy.org/) was employed (see 4.6.2 Database Development).

4.6.4 PepTracker Task Scheduler

A scheduler was implemented on the PepTracker server to run tasks in the background. Highly demanding tasks (CPU and memory intensive) require considerable server resource and time to be processed. Hence, having these tasks running in parallel and competing for resources is not ideal. Instead it is more logical to run these tasks in a queue as background tasks on the PepTracker server. In order to manage such tasks, including data upload and dataset caching, a scheduler was implemented to manage a queue of tasks, execute these tasks and inform the researcher once these tasks were complete. This scheduler is built as a Django application, which the main PepTracker application communicates with. The scheduler supports two types of tasks:

- Naïve tasks: These are ordinary functions, which are wrapped by a very thin execution interface, which in turn is called directly by the scheduler. Whatever is ordinarily returned by the function is cast to a string and returned to the user. If the function raises an exception, the exception message is returned to the researcher, and a full traceback is made available to the system administrator.
- Jobs: These are functions, which are developed specifically for use with the scheduler, as they adhere to a number of conventions and must implement certain behaviours. However, they provide a wider range of functionality compared to naïve tasks alone, including the ability to manually abort the task, a prioritisation scheme for allocating more resources to important tasks, more robust detection of failed tasks, and the ability to return progress report strings to the user while the task is running.

4.6.5 Proteomics Server

The raw data produced by the mass spectrometer is stored by the mass spectrometry facility on a server (known as the Proteomics Server). In order to create a link between datasets and the original raw files, the PepTracker server has an active link with the Proteomics Server. The PepTracker server polls the Proteomics Server for new raw data files associated with samples that have been submitted via the PepTracker application. When found, the PepTracker database is updated to log the location of original raw files, hence maintaining an active link between experimental metadata and raw data.

4.6.6 LDAP Authentication

In order to be more accommodating to researchers, the logging process to PepTracker is linked with the University of Dundee Lightweight Directory Access Protocol (LDAP) server. This means researchers can use the same username and password to access PepTracker as they do for other University IT facilities. PepTracker requests login credentials and sends the username and password to the LDAP server, which confirms whether the details have been entered correctly. Based on the response from the LDAP server, PepTracker can then provide secure access to the application and data. To accommodate non-university researchers, PepTracker also has in-built authentication features.

4.6.7 Graphical User Interfaces

All GUIs in PepTracker have been designed to be user friendly and provide a good user experience of the PepTracker application. The web interfaces in PepTracker are implemented using HyperText Markup Language (HTML) for the general structure, Ajax (Asynchronous Javascript And XML) for dynamic data loading using the REST (REpresentational State Transfer) approach, Javascript to provide advanced functionality and Cascading Styling Sheets (CSS) for styling.

A number of Javascript libraries were used, including JQuery (version 1.3.2), JQueryUI (version 1.7.2), Mootools (version 1.2.1) and YUI (version 2.7.0), to add and enhance the functionality available through the PepTracker interfaces. Features, such as live searching implemented using Mootools, enhance user experience. Furthermore, libraries such as the Google Chart API and Google Visualisation API were utilised to enhance interfaces through display of data in charts and graphs.

The interfaces are made more dynamic using AJAX, which allows the PepTracker interfaces to generate and send requests to the PepTracker Server using HTTP. The PepTracker server processes the requests and returns a response, which can be used by the clients interface to update its display without having to refresh the full interface. The PepTracker web interfaces can be accessed through a web browser, such as Firefox, Safari or IE.

To complement the HTML pages served through Django, the PepTracker application also includes components programmed in Adobe Flex (SDK 3.4). Flex is an open source Software Development Kit (SDK), released by Adobe, for the development and deployment of cross-platform, rich interactive applications that deploy across all major browsers, operating systems and desktops via the Adobe Flash Player and Adobe AIR runtimes. In order to build Flex components the Adobe Flash Builder (formally known as the Adobe Flex Builder) was used. All code in Adobe Flex is written using MXML, an XML-based markup language, for building Graphical User Interfaces (GUIs) and Actionscript for interactivity. This code is built around the Cairngorm framework (see Appendix A. Cairngorm Framework), an open-source MVC-based framework for application architecture. Many of the in-built Adobe Flex display components (Accordions, Menus, Data Grids, Titled Windows, and Buttons) data types, effects (Zoom, Blur, and Dissolve), and charts were used and/or extended to generate the PepTracker interfaces.

The Adobe Flex code base is used and extended within Adobe AIR (Adobe Integrated Runtime). Adobe Air provides a cross-platform runtime environment for building rich interactive applications that can be deployed as desktop applications. This provides huge benefit for a developer, as code for a web client and desktop client only needs to be written once, with minor adjustments the same code can be deployed on both web and desktop platforms. Adobe Air was used to create and deploy a PepTracker desktop client that is downloadable through the PepTracker Server.

4.6.8 Protein Identification Strategies

There are two main approaches to protein identification using mass spectrometry, the first approach analyses intact proteins (top-down, protein-centric) whereas the second approach involves digesting proteins into peptides and attempting to identify these peptides (bottom-up, peptide-centric). With a peptide-centric approach, more

complex computation is needed to reassemble the identified peptides to the corresponding proteins.

A protein-centric approach often makes it easier to map between identified proteins and a search database, as the analysis algorithms are working with a full protein identified via MS. Furthermore, this approach negates the need for the time-consuming protein digestions required with a peptide-centric approach. However, in order to identify these large proteins, highly sensitive MS equipment is needed. Average mass spectrometers find it difficult to handle large proteins (>50kDa), particularly due to the challenges faced in chromatographically separating proteins. Often the resultant mass spectra for a protein will show numerous peaks due to the stochastic changing nature of separation and will be further complicated due to multiply charged proteins. Furthermore, many proteins can be lost in the initial experimental protein separation phase. These challenges limit the protein-centric approach to the analyses of isolated proteins or simple protein mixtures at best.

However a protein-centric approach does have advantages when attempting to identify PTMs. Having the ability to identify and measure a whole protein allows for data analysis algorithms to determine which modifications may occur in conjunction with one another, whereas in a peptide-centric approach, modifications identified on separate peptides from the same protein may come from different pools of the same protein.

Biologists analysing proteomes often favour the mature, bottom-up approach, due to its ability to handle larger proteins and more complex samples. This is the strategy used by the Lamond Laboratory. With a peptide-centric approach, biologists must first digest their sample proteins to peptides using an enzyme. The peptides are then passed through the mass spectrometer to produce mass spectra (MS) and can be further fragmented within the mass spectrometer (MS/MS). The data analysis algorithms used to determine the proteins in the constituent sample are complex. Measurements from the mass spectrometer, which actually measures ions, must be mapped to theoretical peptide masses, calculated from a proteomics database, before finally being reconstructed to proteins.

An additional complexity that must be considered is the relationship between proteins and peptides. In fact, a peptide can belong to multiple proteins due to sequence homology. This sharing of peptides makes it increasingly difficult for a search and matching algorithm to determine exactly which protein was detected. Hence, software, such as MaxQuant, will report protein groups rather than single protein identifications to deal with this ambiguity.

With a peptide-centric approach, only a fraction of the total peptide population of a protein will be identified, therefore, information about a protein sequence is lost. This loss of information combined with the many-to-many relationship between proteins and peptides is particularly problematic when attempting to detect protein isoforms, which may differ by only a short string of amino acids. These amino acids might be captured in one peptide, which requires the mass spectrometer to identify this single peptide to accurately determine which protein isoform was detected. However, the stochastic nature of MS makes it very difficult to identify a specific single peptide.

When using the peptide-centric approach, data analysis algorithms must take into account the number of peptides identified and the sequence coverage of a protein to accurately convey the confidence in protein identifications. This raises challenges for the identification of low abundance proteins for which there will be fewer peptides available for detection in mass spectra that is dominated by high abundance species, hence reducing the likelihood of detection.

Furthermore, it is also challenging to identify PTMs using the peptide-centric approach, as it relies on the detection of peptides with the modification to enable accurate inferences. To aid with PTM detection, biologists will often turn to experimental protocols that enable them to enrich their sample for certain types of PTM. Another possible technique that could be employed to detect PTMs from mass spectra is a spectrum-centric approach, whereby the ion fragmentation spectra could be used to build a spectral library. This spectral library would capture previously identified and validated spectra, which in turn could be used by a search and match algorithm to identify the same peptide modifications in new spectra. By creating spectral libraries, it would be possible to benefit from previously found and annotated spectra. This may also help in situations where there are overlapping spectra of two separate peptides. A

data analysis algorithm could distinguish between these peptides by matching multiple, validated spectra to the results.

PepTracker takes data generated in the Lamond Laboratory using the peptide-centric approach and aims to provide visualisations, such as the Protein-Peptide Alignment Map (see Figure 24) to aid in the analysis and evaluation of protein identifications. Furthermore, during this thesis, additional tools and techniques were developed to help combat some of the challenges with a peptide centric approach, such as the detection of novel isoforms. Chapter 7: Protein Isoform, Localisation and Turnover Analysis, describes strategies that deal with missing protein sequence data by making use of protein properties to determine novel isoforms or pools of proteins.

4.6.9 Protein Definitions

A commonly understood problem in proteomics is related to the definition of proteins. In the proteomics field there are a number of major databases that define proteins using their own specific methods, for example some databases may allocate protein isoforms each with their own identifier whereas other databases may chose to group these isoforms together. This results in each database using a customised convention for assigning protein identifiers. Furthermore, the identifiers used can be unstable and change between database versions, being created, deleted, updated or merged. Translating between these identifiers remains a major challenge.

When choosing the identifier mechanism to be used for the software created during this project, a number of issues were considered, including the popularity of the database, the longevity of the database and its developers and the most relevant database to the area of quantitative proteomics addressed during this work. It was important to choose a single identifier to be used to map proteins within datasets, as this enables the comparison and cross analysis of datasets. Furthermore, it was vital to choose a database that is robust and has a significant development effort supporting it, both in terms of funding and developers.

For the software within this project the UniProt Knowledgebase Identifier was selected. This decision was made due to UniProtKB positioning amongst the other database vendors, relevance and usefulness. UniProtKB is a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and

the Protein Information Resource (PIR). UniProtKB combines both automated protein identification with manually curated protein specifications. It holds these data in two subsections, UniProtKB/Swiss-Prot, containing data from manual analysis of literature and reviewed computational analysis, and UniProtKB/TrEMBL, containing computationally annotated entries that are reviewed in due course and transferred to UniProtKB/Swiss-Prot. In addition, UniProtKB collects a range of data from various sources, including biological ontologies, classifications, cross-references and evidence of annotation. This information supplements the basic protein information, such as protein sequence, name and description, to provide researchers with rich annotation that can aid with functional analysis of proteins. UniProtKb releases four weekly builds that are well annotated with version information for each entry in the database. Furthermore, UniProtKB appears to have a dominant position due to its takeover of the International Protein Index database (last release 27th September 2011), another popular protein cross-reference database curated by the European Bioinformatics Institute (EBI).

4.7 Data Security & Quality Control

The system has been designed securely, with restricted researcher access. All University of Dundee (UoD) researchers have the option of Lightweight Directory Access Protocol (LDAP) access, which negates the use of an additional username and password. Instead researchers login with their UoD credentials, which are authenticated against a UoD LDAP Server by PepTracker. Having multiple login credentials is inconvenient for researchers and often results in researchers writing down passwords or sharing them. Furthermore, the UoD policy actively encourages researchers to routinely update their password, hence increasing security.

However, some users, such as external collaborators, are not directly affiliated with the UoD, hence do not have UoD access credentials. In order to accommodate external researchers, PepTracker implements its own authentication that allows researchers to create a customised PepTracker username and password for use with the system.

Access is restricted due to privacy concerns for data. Within PepTracker each researcher belongs to single or multiple groups. Each group typically represents one laboratory. PepTracker researchers have access to all metadata and quantitative MS data uploaded to the PepTracker database by researchers within their group. In

addition to this read access, researchers are provided with editing privileges to the MS submissions and linked data that they themselves have generated or been allocated permission to edit by the owner of the data. Due to this open data access policy, direct access to PepTracker is controlled via usernames and passwords. Data that may be useful to external researchers is shared through externally available PepTracker tools, such as the PFL and Turnover Viewers. These external tools require registration, which is free of charge, but is essential as it allows monitoring of demand for these tools.

Furthermore, an important aspect of controlling access to PepTracker revolves around data quality issues. The usefulness of PepTracker tools relies on data being generated in a known style, well and consistently annotated and of high quality and adequate stringency. Due to the variance in protocols and instruments, the quality of data generated can vary. One of the future goals of PepTracker will be automated evaluation of data quality at the point of data upload. This could be accomplished through using existing PepTracker datasets as a baseline to generate an overall profile of what is expected in different types of dataset. These profiles could be compared with new datasets to highlight inconsistencies and/or unexpected inputs that could suggest data quality issues. This would open up the possibility of external researchers being able to contribute data that could add to the usefulness of PepTracker tools.

4.8 Discussion

The PepTracker data environment (http://peptracker.com) is a scalable, fast and secure, large repository for proteomics data, accessible via the web. This proteomics repository has huge potential and is already growing rapidly based on current usage. The PepTracker software has three main components, MsTrack, DataVault and ProteinLibrary. These components work together to provide an integrated data environment for mass spectrometry data management and analysis.

The MsTrack component implements a LIMS that tracks all data produced and analysed in the laboratory. Data entry is standardised and where possible drop down menus are used for entering metadata associated with each experiment, ensuring that all experiments are annotated consistently and in detail. New development efforts have also been allocated to creating an online laboratory book, LabTracker (see Appendix E). This software aims to provide computerised data collection at the lab bench via an iPad, as well as providing researchers with easy-access to sets of common

tools, such as molecular calculators, molecule viewers, reagent database searches etc. In the future LabTracker be further integrated with PepTracker so that each MS submission is also annotated with a protocol from a users online laboratory book.

Having thorough metadata recording is vital for current and future data analytics. A huge range of biological questions can be answered by taking measurements made on the levels and properties of every protein in every cell type and under a wide range of experimental conditions, thanks to the detailed annotation with associated metadata. Through accurate tagging and aggregation of complex quantitative biological data there has been development of pioneering new approaches that allow multidimensional analysis to be carried out, suitable for benefitting both basic and applied biomedical projects. The Lamond Laboratory is already generating vast amounts of data, which can be used along with the metadata recorded as a test bed for mining.

After logging samples into the LIMS and submitting them to the mass spectrometry facility, researchers can continue with other work-related activities while the workflow automatically deals with logging the data files from the mass spectrometer and generating a notification email when all samples have been run.

Currently researchers download files and run them through third party software, e.g. MaxQuant. However, it is envisioned that this step could be automated in the future, so that the PepTracker system initiates the MaxQuant analysis (or equivalent functionality), before automatically loading the results into the data repository. Once the quantitative results have been uploaded to the PepTracker system, researchers can use the DataVault component of PepTracker to view automated visualisations of their data, such as charts and network maps, and to interact with their data. Data interaction has been considered as being of high importance as researchers working on proteomics data must find a way of combating the complexity of the data. By focusing on interaction, PepTracker provides researchers with the ability to explore their results and drill down into their data and hence convert the many data points to meaningful biological insight. In feedback sessions, researchers have commented positively on the usefulness of interaction-focused visualisation, which allows them to make sense of their data much more rapidly compared with previous attempts using other software.

Finally the ProteinLibrary component maximises on the benefit of having datasets stored in a meaningful structure by providing researchers with the ability to query datasets. The output search results of the querying are presented with the aid of additional visualisations, such as protein peptide alignment maps, which deliver the information to be conveyed in an easy human understandable form that is more appealing than traditional text or table methods. Furthermore, the ProteinLibrary allows researchers to store protein groups of interest. These protein groups can be generated in a number of ways and used in the analysis of MS datasets.

The Lamond Laboratory experience of data analysis is that analysis and management of MS data (rather than experimental design or data generation) has been rate limiting for all proteomics work. Therefore, the creation of PepTracker has already been transformational for enhancing proteomics projects and is integral to the future of MS experiments.

Whilst creating PepTracker, the challenges with existing software were considered carefully (see 1.1.4 Challenges in the Proteomics Field). In order to ensure that PepTracker meets user requirements and is taken up and utilised by researchers a user-centred approach was taken whereby there was close communication between software developer and biologists. This is seen as an essential part of the creation and future development of PepTracker. Having computer scientists working directly within the Lamond Laboratory and in daily contact with the molecular and cell biologists, performing the MS studies, results in immediate feedback and suggestions for new features and improvements. Thus, the project has been led by the biology and the needs of the researchers and a major emphasis was placed on the ease of use and intuitive design of researcher interfaces.

Furthermore, to support the software, especially new users, a number of video and written tutorials have been implemented. In addition, a mailing list was created to which researchers can report errors, ask questions and provide feedback. The PepTracker developer aimed to respond to user requests/queries as soon as possible. Furthermore, over the three-year development of the PepTracker software there were an additional seven developers, both at undergraduate and graduate levels, involved in the development. This ensured PepTracker was created in an extensible manner where multiple developers can contribute to the software. Furthermore, having

several developers contribute promotes security for the software beyond this PhD project, as the software is not reliant on only one developer. This ensures continuity for the project and continued support for users, which are both attributes of successful software developments by research laboratories. Beyond this work, the continued success of PepTracker is reliant upon sustaining a team of developers who are able to maintain and support the software.

In addition, to ensure that users can access the PepTracker software without high licensing fees it was decided to make the PepTracker software freeware and ensure its implementation is based on open source technologies, such as Apache CentOS Server, Python server-side code and PostgreSQL database engine. This means users have no additional cost in implementing the PepTracker solution. In the future, the PepTracker codebase could also be released as open source, which would further encourage the community to actively contribute to the development of the PepTracker suite.

PepTracker addresses the main issues found with other software targeted at the downstream analysis of quantitative proteomics data (see Table 3: Comparison of Major Mass Spectrometry Data Analysis Software.). There is an evident gap in software designed for mass spectrometry analysis, whereby no available current software provides central data warehousing functionality. However, this functionality is built into PepTracker, and is viewed as providing a key advantage for data analysis possibilities. Furthermore, the software reviewed lack in functionality to capture, organise and store metadata in a systematic manner. Only a select few programs had a built in LIMS system that can systematically record metadata and link this additional metadata to the resultant data files. Within PepTracker the LIMS functionality is integrated with automated mass spectrometry submissions, which guarantees metadata recording, and much attention has been placed in ensuring metadata capture is carried out efficiently as possible. Furthermore, PepTracker is able to deal with both label-free and labelled experiments, which is vital functionality missing from other software.

The PepTracker environment is already in daily use and in active development within the Lamond Laboratory. More recently, users from other laboratories within the College of Life Sciences have been given access to the core PepTracker functionality.

This allows for evaluation with researchers working in external laboratories with modified protocols and experimental setups.

Moving forward with the project it is intended that the PepTracker resource will be expanded in terms of its development and support a larger group of researchers. Plans include widening access to the other groups in the Wellcome Trust Centre for Gene Regulation and Expression, as well as other researchers in the College of Life Sciences at the University of Dundee. It is envisioned that this could take place by maintaining a single data warehouse within the College of Life Sciences but sharding the web application across a number of servers to deal with the higher demand of users. To support external collaborators it is intended that the server will be packaged up into a downloadable installation. External users would be provided with a database server setup for the main data warehouse and a web server setup for the PepTracker tools. A number of configuration interfaces will be created to help users with external setups get started and populate the basic lookup tables. Initially a few test groups, external to the University of Dundee, will be selected to allow evaluation and trouble-shooting of practical issues involved in distributing the software and making it widely accessible.

The collection of uploaded datasets in PepTracker provides an ideal target population that is large enough to be used as a baseline for identifying patterns and trends. As described in 1.4.1 Super-Experiment Data Analysis, a ‘super-experiment’ can be understood as analysis involving multiple independent datasets in which each dataset provides value to the analysis of every other dataset in the collection. An example of what a ‘super-experiment’ can yield is demonstrated by the Protein Frequency Library (PFL), which is described in Chapter 5: Multidimensional Analysis with IP Experiments. The PFL study demonstrates the huge added value of integrating and utilising all data collected in every pull-down experiment.

It is expected that the set of tools for super-experiment analysis will continue to be expanded and interfaces will be refined and updated to make it easier and faster to analyse and compare datasets. For example, new tools will be added for the visualisation and analysis of PTMs and to combine analysis of PTM data with Spatial Proteomics approaches. Tools will be provided for researchers to create and annotate libraries of datasets, based either on data outputs from experiments, or from online resources, e.g. according to GO annotation terms, and to use these datasets flexibly for

comparative analyses and statistical comparisons. It is also hypothesised that this will spark collaboration where datasets from other researchers could be imported into PepTracker to enhance the analyses.

Chapter 5: Multidimensional Analysis with IP Experiments

5.1 Summary

The reliable identification of protein interaction partners and how such interactions change in response to physiological or pathological perturbations is a key goal in most areas of cell biology. In order to identify protein-protein interactions pull-down experiments are carried out, whereby proteins, and their interacting partners, are isolated from experimental samples and then run through the mass spectrometer. This method allows biologists to identify the proteins that are interacting with a specific protein of choice under certain experimental conditions.

Typically, for a pull-down experiment, the mass spectrometer can identify 500 or more proteins. All of these do not constitute genuine interaction partners and most of these will be contaminants that are introduced either from the environment or during the experimental procedure. In order to distinguish the real interaction partners from these contaminants a number of methods can be employed, including increasing the stringency of the experimental procedure or repeating experiments multiple times to have many controls. However, these methods risk biologists losing low abundance or low affinity real interaction proteins and repeating experiments is very costly and time intensive. An alternative, more beneficial approach involves identifying all proteins (both interaction partners and contaminants) and then finding a method of annotating them accordingly. This requires an accurate and reliable mechanism of distinguishing between the two categories of proteins. This method can be based on how a biologist would carry out this task manually using their experience to make informed judgements.

The aim of this work was to capture this process in an automated system that makes use of previous experimental experience through the datasets generated. Due to the complexity of the cell biology system and the human aspect of doing this task reliably, it would be better for such a system to annotate proteins with a likelihood of being a real interaction partner rather than simply categorise as contaminants and non-contaminants. Using these predictions, the likely contaminants can be used to re-normalise a dataset that may have been affected by experimental abnormalities.

The resultant tool was the Protein Frequency Library (PFL) (Boulon et al., 2010a). The PFL collates a large amount of experimental data and generates automated annotations for proteins based on their frequency of detection in experiments. These annotations can be further improved by the PFL through creating a customised PFL that generates annotations using only experiments that meet certain experimental conditions. In order to achieve this technically, methods from the world of BI were employed. BI allows for fast train-of-thought analysis through rapid responses, making it an ideal tool for enabling biologists to carry out this type of complex analysis. BI has rarely been used in the field of proteomics, however this case study validated the usefulness of the analytical benefits provided by BI techniques.

To enable biologists to carry out fast and effective analysis, the PFL functionality was integrated into PepTracker. Within PepTracker, researchers can filter a selected dataset using customised PFL frequencies and also normalise an entire dataset using the likely contaminants identified by the PFL frequencies. Furthermore, to enable external researchers to benefit from the PFL in the analysis of their own data, a custom interface was built. This interface was partly developed by an undergraduate honours project student whom I supervised. The interface provides an intuitive method for researchers around the globe to access protein annotations, coming from an extensive data repository of high quality and high consistency data.

Chapter 5 describes the development of an analysis methodology for IP experiments, focusing first on the purpose of IP experiments and established analysis techniques (section 5.2), following with a description of the PFL (section 5.3), its implementation (sections 5.4) and finally a discussion on the use of the PFL (section 5.5).

5.2 Background

Pull-down experiments are used to identify proteins that interact with one another. To create a sample in a pull-down experiment, a protein is isolated from a mixture along with its interacting proteins. In order to isolate these proteins from the mixture, an affinity matrix or ‘bead’ is used. Protein identification of the sample yields the protein of interest along with the interacting proteins. The high sensitivity of MS technology has increased the total number of proteins identified in each pull-down experiment. However, the majority of these proteins usually represent contaminants, including proteins that bind non-specifically to the affinity matrix. Thus, despite many technical

improvements made in recent years, the unambiguous discrimination between genuine protein interaction partners, either stable or transient, and contaminants, remains one of the major challenges in the field.

Most researchers have sought to identify specific protein interactors by reducing or eliminating the background of non-specific proteins, either through biochemical or data analysis strategies. For example, at the experimental level, the buffer stringency can be increased to reduce binding of low affinity contaminants, and a 2-step tandem affinity purification method can be used rather than a one step procedure (Babu et al., 2009, Rigaut et al., 1999). However, this can decrease the yield of proteins recovered and risks losing low abundance and/or lower affinity specific protein interaction partners. Alternatively, on the data analysis level, several approaches have been used to identify and thereby discard the putative contaminants that are recovered after purification. For example, bioinformatics can be employed to measure “confidence scores”, by comparing the results of interaction studies with either predicted protein-protein interaction data, or with previous results described in literature (Blow, 2009), or by integrating different properties of the interaction network generated by the analysis, e.g. interaction bidirectionality etc. (Cloutier et al., 2009, Ewing et al., 2007).

The combination of quantitative MS and differential labelling of proteins with heavy isotopes, especially SILAC (Ong et al., 2002, Ong and Mann, 2006), can also help to distinguish between specific and non-specific binding proteins in a co-immunoprecipitation (co-IP) experiment. This is achieved through the inclusion of an internal negative control, which allows for direct comparison between the relative levels of each protein present in the control and experimental samples (see Figure 27).

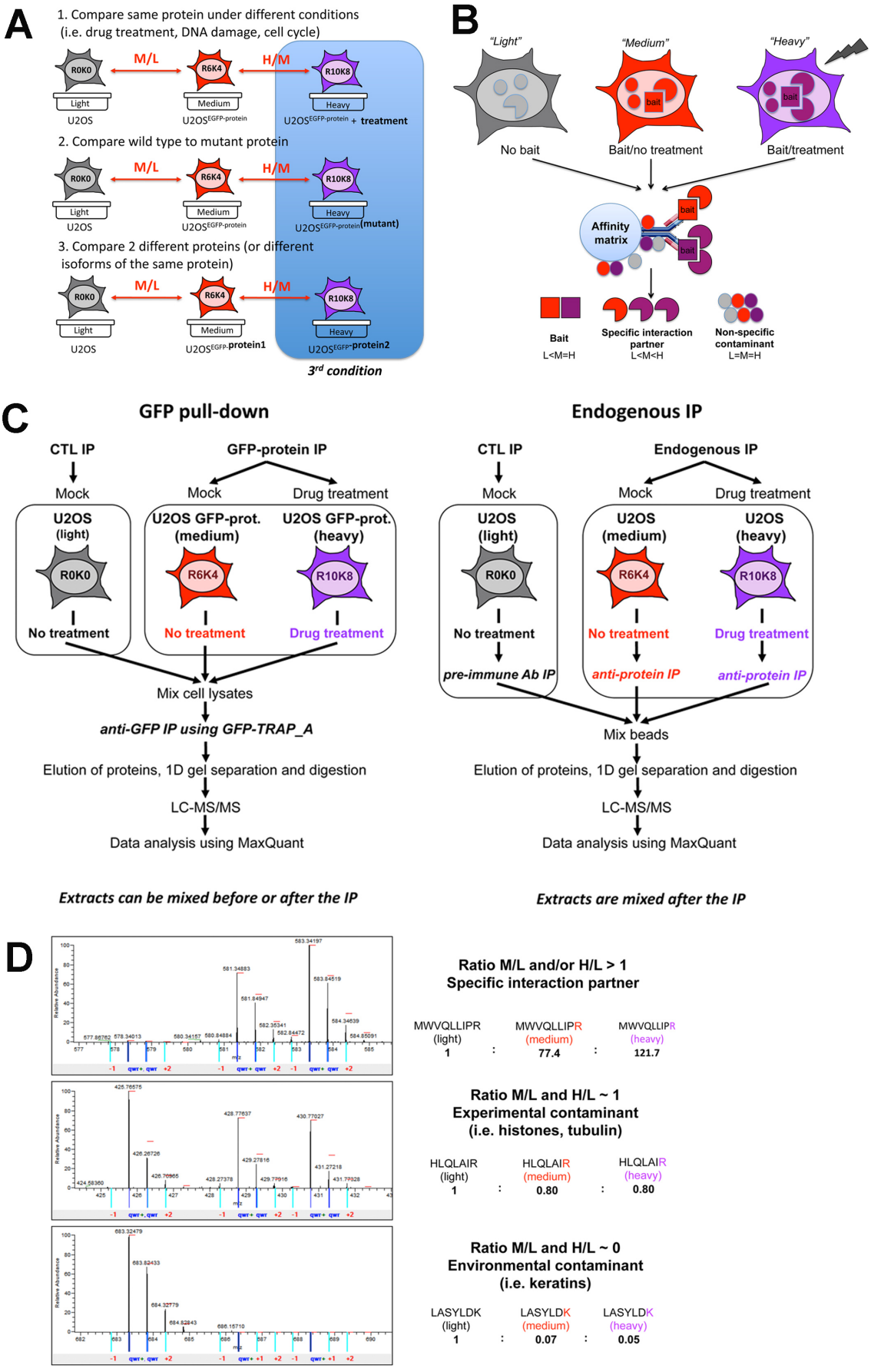


Figure 27: Overview of triple SILAC-based analysis of protein interaction partners.

(A) Metabolic labelling of cells in culture using the triple SILAC approach can be used to detect specific protein interaction partners and dynamic changes in protein interactions under different biological conditions. Examples include comparing control conditions with; (i) treatment with chemical inhibitors/stress etc. (ii) effect of mutations in the bait protein or (iii) isoform-specific interactions. "Light" media refers to normal environmental isotopes of carbon, nitrogen and hydrogen, i.e. "unlabeled" ^{12}C, ^{14}N and ^{1}H, while "medium" and "heavy" media refer to cells grown in medium containing heavy isotope-labeled arginine (R) and lysine (K) as follows; medium - $^{13}C_6$-arginine (R6), 4,4,5,5-D_4-lysine (K4), heavy - $^{13}C_6{}^{15}N_4$-arginine (R10), $^{13}C_6{}^{15}N_2$- lysine (K8). (B) Diagram illustrating SILAC principle of differential labeling and how specific interacting proteins have higher ratios of heavy isotope- labeled peptides as compared with non-specific contaminants. (C) Overview showing workflow in a representative triple SILAC analysis of protein interactions and their response to inhibitor treatment for either GFP-tagged or endogenous cell proteins. (D) Example of MS spectra for representative peptides illustrating either a specific protein interaction partner (top), an internal contaminant binding non-specifically to the beads (middle) and an external environmental contaminant, e.g. keratins (bottom).

SILAC thus objectively identifies proteins that can bind non-specifically, e.g. to the affinity matrix and/or the fusion tag, and by comparison highlights proteins that bind specifically to the bait protein (reviewed in (Ranish et al., 2007, Vermeulen et al., 2008)). The Lamond Laboratory and others have used this isotope-based, quantitative MS approach to characterise both tagged and endogenous protein complexes in mammalian cells (Blagoev et al., 2003, Selbach and Mann, 2006, Trinkle-Mulcahy et al., 2006, Trinkle-Mulcahy et al., 2008a). Related differential isotope-based labelling strategies, combined with MS, have also been used to analyse specific binding proteins (Brand et al., 2004, Tackett et al., 2005).

However, relying upon isotope labelling ratios alone does not entirely solve the contaminant problem. Indeed, it is often impossible to establish a threshold ratio level in these experiments that eliminates all of the contaminating proteins without discarding, en passant, genuine interaction partners of lower abundance and/or lower binding affinity. This issue was previously addressed by systematically identifying proteins that frequently occur in pull-down experiments. These proteins were documented in a "bead proteome", which provided a filter to help discriminate between specific interaction partners and the inevitable non-specific background (Trinkle-Mulcahy et al., 2008a). The bead proteome has proved a useful tool for understanding contaminant behaviour and how it can be modelled. It has been used extensively in the analysis of many pull-down experiments to guide the identification of specific interaction partners. However, a number of limitations have been identified

of the static list of proteins included in this bead proteome. It has increasingly become apparent that this list can never be definitive due to the constant changing nature of protein discoveries. The original bead proteome paper focused on protein clustering and frequency, as the method used to determine the proteins belonging to the bead proteome, however, this is an unreliable, stand-alone method of determining contaminants. The original paper identified these limitations:

> *"The bead proteome filters thus provide a useful and objective resource that can be consulted by cell biologists to help avoid expending time and effort on the analysis of proteins that may prove to be simple contaminants. In the future,* ***accumulating information from many laboratories on the range of nonspecific protein interactions observed using different cell types, extracts, tags, and affinity matrices*** *will provide an invaluable resource and we propose* ***this should be established as a freely accessible online database****." (Trinkle-Mulcahy et al., 2008a).*

Clusters of contaminants are not always well defined in datasets, nor are all the proteins found within the clusters, guaranteed to be contaminants. Furthermore, it is impossible to generate an absolute list that can apply to all pull-down experiments, as the contaminant proteins differ based on experimental conditions. Variations in experimental design, such as cell type, cell extract, organism, beads and sensitivity of the mass spectrometry machine, have all been shown to have an impact on the contaminants found in pull-down experiments. Rather than a static list of contaminants, a more objective approach would involve a dynamic list of contaminants that can be customised for individual experiments.

This chapter presents a methodology for the reliable identification of specific protein interaction partners that overcomes limitations with the previous bead proteome approach. This methodology draws on data analysis strategies from the field of Business Intelligence (BI) and applies them to integrate complex datasets arising from MS pull-down experiments. This methodology is used to generate a Protein Frequency Library (PFL) that can be customized to the conditions of specific experiments and continually updated. The PFL can be used as a specificity filter to discriminate specific protein interactions and as a tool to normalise datasets and hence facilitate comparison of separate experiments.

5.3 Results

The analysis of pull-down data theoretically allows for (i) the discrimination between contaminants and genuine interaction partners, and (ii) the characterisation of changes in protein complexes under specific biological conditions.

5.3.1 Discriminating Specific from Non-Specific Interaction Partners

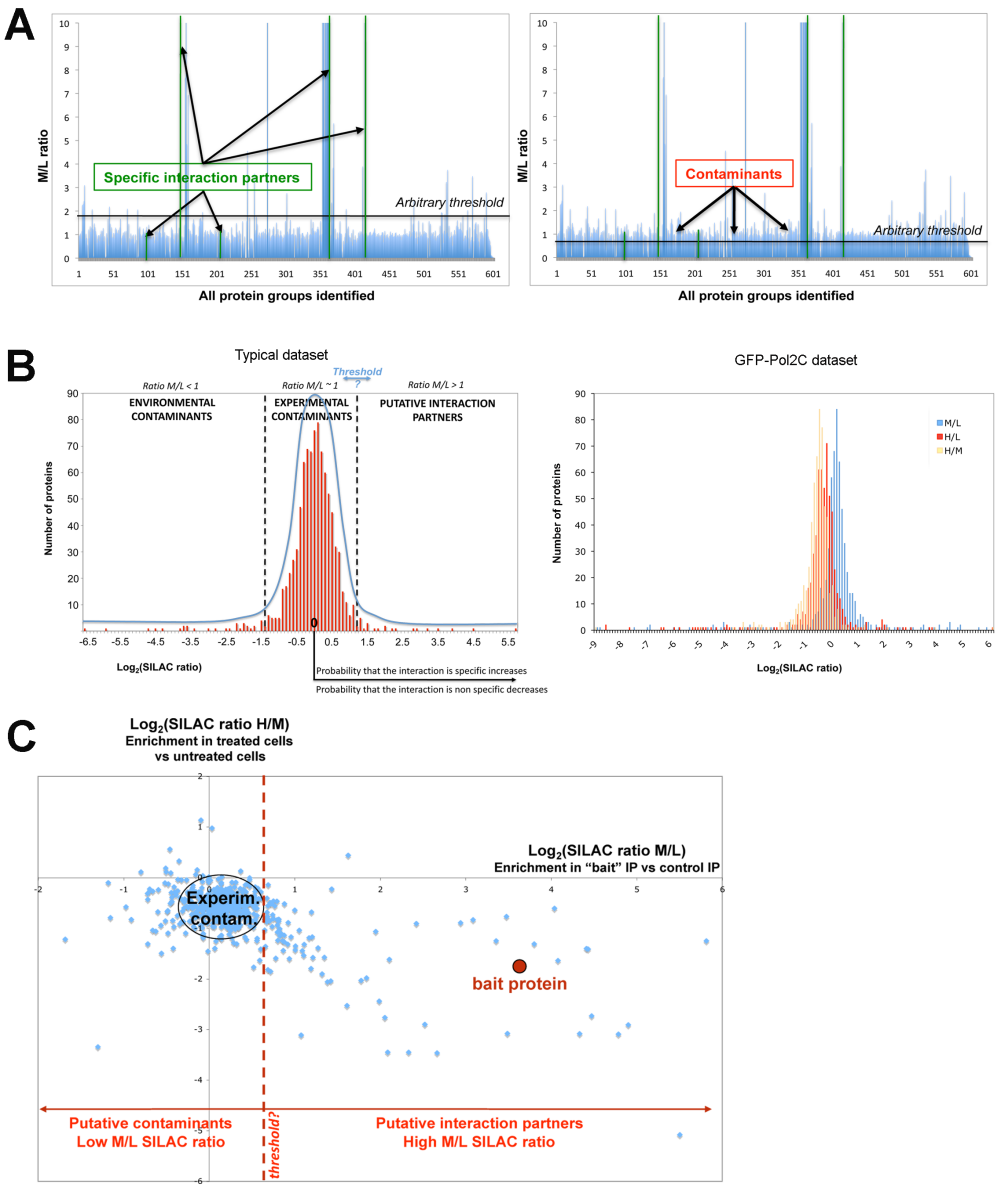


Figure 28: Visualisation of contaminant profiles and threshold levels.

A representative example of a triple SILAC co-IP experiment, using GFP-Pol2C as bait in cells either with, or without, α-amanitin treatment (Boulon et al., 2010b), was used to generate the graphs shown. (A) Graphs showing median SILAC ratios for every protein group identified and quantified by MaxQuant (604 distinct protein groups), with each protein group plotted on the X-axis and the median SILAC value for that protein group plotted on the Y-axis. Two arbitrarily chosen thresholds are illustrated (black horizontal lines in left and right panels). (B) Representative ratio distribution plots. Data are plotted as a histogram with log₂ SILAC ratios on the X-axis and number of proteins for a

given ratio on the Y-axis. Non-specific contaminants reproducibly cluster in a Gaussian (normal) distribution centered ~zero (left panel), although the exact mean can deviate from zero due to experimental variability, as seen for GFP- Pol2C dataset (right panel). (C) Data from the GFP-Pol2C dataset plotted with log_2 (M/L) SILAC ratio on the X-axis and log_2 (H/M) SILAC ratio on the Y-axis, with each point corresponding to the ratio value for a specific protein group. The bait protein is shown in red. Putative experimental contaminants cluster around the origin.

Figure 28 shows an example of data analysis from a representative SILAC pull-down experiment in which a tagged complex (Pol2C) was affinity purified from U2OS cells. In this case, after in-gel digestion, the MS analysis identified and quantitated over 4,000 peptides that were assigned by MaxQuant to 604 human protein groups. For each protein group (X-axis), a median M/L SILAC ratio was calculated from all of the individual peptide values determined and shown plotted on the Y-axis (see Figure 28A). This shows that a minor group of proteins have a high SILAC M/L ratio (>2), while approximately 80% of the proteins (i.e. over 480 out of a total of 604 protein groups) have a SILAC ratio <1.4. As described above, the former are strong candidates to be specific interaction partners (see Figure 28A, green columns), while the latter are more likely to be non-specific interaction partners. However, experience has shown that some bona fide specific interaction partners can have SILAC ratios lower than abundant contaminants (e.g. in the range ~0.6-1.4). Thus, when setting the threshold to an arbitrary value it is important to understand that if the selected threshold is high, although most or all contaminants will be eliminated, low abundance and/or low affinity genuine interaction partners will be lost. Conversely, if the threshold value is chosen to be low, with the aim of identifying all low abundance and/or low affinity partners, a larger number of contaminants will remain (see Figure 28A, comparison between left and right panels). It is therefore not possible to use a specific ratio value as a threshold that consistently and unambiguously separates the specific from the non-specific interaction partners.

Another way of visualizing the same SILAC pull-down data is to plot the ratio distribution as a histogram. Thus, for either M/L, H/L or H/M SILAC ratios, the number of proteins with each ratio value is plotted on the Y-axis, against log_2 SILAC ratio values on the X-axis (see Figure 28B). Here, non-specific, experimental contaminants reproducibly cluster in a Gaussian (normal) distribution centred at the log_2 ratio ~0 (which corresponds to a SILAC ratio ~1) (see Figure 28B left panel). Theoretically, the

normal distribution should be centred on a $\log_2$ value of exactly zero but in practice this varies between individual experiments and the actual mean can be either higher or lower, even for the separate M/L, H/L and H/M ratios measured within a single triple SILAC experiment (see Figure 28B, right panel). In contrast, putative interaction partners are expected to show $\log_2$ ratio values greater than the mean of the Gaussian curve, while environmental (external) contaminants always have values lower than the central mean value (see Figure 28B, left panel). The Gaussian curve can be useful to help refine the analysis of predicted specific interacting proteins, using a mathematical description of the protein distribution. However, there is still no single ratio value to reliably distinguish specific from non- specific proteins.

Figure 28C shows a third way of visualising the data, i.e. by plotting $\log_2$(H/M) (Y-axis) versus $\log_2$(M/L) (X-axis) SILAC ratio values for all proteins identified in the triple SILAC co-IP experiment using GFP-Pol2C as bait. This visualisation provides an indication of both the specificity of the interaction (M/L ratio), and the changes occurring between the two conditions tested (H/M ratio). From this graph it is evident that most proteins have SILAC ratio values that cluster around the origin (see Figure 28C, circled proteins). As these proteins have $\log_2$(M/L) and $\log_2$(H/M) ratios of approximately zero, they have a high probability of being contaminants. Due to small variations in each experiment, e.g. volume differences when mixing extracts, the contaminants typically cluster around values that can however deviate from zero (see Figure 28, B, right panel, and C). In contrast, putative specific interaction partners are present in the right side of the graph. But as described above, and regardless of how the SILAC pull-down data are visualised, the problem remains that a significant overlap invariably exists between the SILAC ratio values of specific interaction partners and contaminating background proteins. Thus, although the SILAC approach is a powerful approach to identify stable interaction partners, it is observed that relying upon SILAC ratios alone is often not enough to reliably identify bona fide interaction partners of lower abundance and/or lower binding affinity. To address this problem, additional objective criterion was sought to add to the analysis. Thus a strategy was developed based upon systematically annotating each protein in the proteome with its frequency of detection in a database of independent co-IP experiments, creating what is termed a Protein Frequency Library (PFL). Hence, the PFL provides a probability estimate for each

protein to be a contaminant, which is independent of the information given by SILAC ratios and, therefore, can be applied to analyse both SILAC and label-free data.

5.3.2 Sun Model and the Protein Frequency Library

In order to quantify the analytical requirements that typify quantitative SILAC pull-down experiments, a logical model was constructed. This takes form as a "Sun Model" (see Figure 37), which shows the "measures" (e.g. SILAC ratio values, number of peptides identified etc.) in the centre of the diagram and the "dimensions" (e.g. type of affinity matrix, type of extract, date, user etc.) radiating from the centre. The hierarchies that can exist within each dimension (e.g. date can include, year, month, day etc.) are symbolised by the levels marked along a dimension line. This logical model was then converted to an OLAP cube implementation. The PFL was extracted using the logical model combined with the OLAP cube, focusing on the measure called "Identified".

The "Identified" measure signifies whether a given protein was identified and quantified in a particular co-IP experiment that is currently in the data repository. The data repository used in the initial tests contained 38 SILAC co-IP experiments, but it is important to note that the PFL can also be generated using data from non-SILAC, label-free experiments. Proteins in the initial PFL were identified via IPI accession number, which provides a comprehensive description that is consistent with the output from MaxQuant. It was noted that, due to the continuously updating IPI identifiers, proteins were mapped to the most current identifier and, thus, multiple occurrences of the same protein accession number, in a single experiment, were only allocated the weighting of a single identification and quantification in the frequency value of any generated protein library annotation. Using the "Identified" measure, the number of times each protein appeared in all 38 experiments in the database was calculated, giving rise to a deduced 'frequency of detection' for each of the 10,623 IPI numbers described by the datasets. This value was used in the generation of the initial PFL.

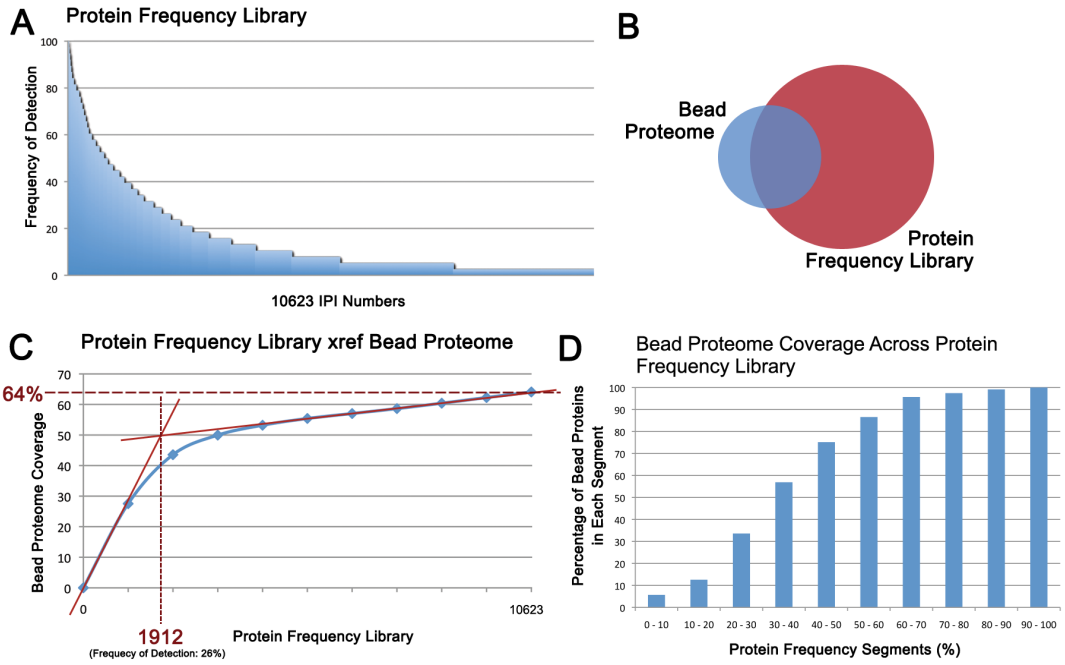


Figure 29: Protein Frequency Library construction and validation.

(A) The sun diagram was used in conjunction with an OLAP cube to analyse the frequency of protein detection in a database containing data from 38 separate SILAC co-IP experiments. Graph illustrates the frequency of detection (Y-axis) for 10,623 separate IPI numbers (X-axis). This defines a Protein Frequency Library (PFL). (B) Comparison of data from the current PFL and a previously determined list of "bead proteome" contaminants (Trinkle-Mulcahy et al., 2008b). (C) Correlation between the "bead proteome" coverage (Trinkle-Mulcahy et al., 2008b) and the PFL, with PFL proteins ranked from highest to lowest detection frequency (left to right). (D) Comparison, for each 10% PFL segment, measuring the number of "bead proteome" proteins (Trinkle-Mulcahy et al., 2008b) found in that segment versus the total number of proteins found in that segment.

The PFL graph presented in Figure 29A shows a visualisation of the frequency of detection plotted against all proteins that were identified and quantified in any of the 38 experiments. In this graph, each protein is shown sorted from the highest to the lowest percentage. Hence, the proteins appearing nearest the origin of the graph have the highest probability of being contaminants.

The PFL was compared with the previously characterised "bead proteome", which contains 3,400 separate human IPI numbers that were frequently found in 27 independent SILAC pull-down experiments (Trinkle-Mulcahy et al., 2008b). The bead proteome includes many abundant factors, such as histones, cytoskeleton and heat shock proteins, and was thus extrapolated to include most members of these large protein families. Although they all potentially can behave as common contaminants,

not all are either expressed, or detected, in every cell type or pull down experiment. An overlap of 64% was observed between the static bead proteome and the PFL, as shown in the Venn diagram (see Figure 29B). The 36% of bead proteome proteins that were not present in the PFL were mostly additional members of large protein families that did not appear in this set of 38 pull-down experiments.

Further comparison shows that most of the common proteins listed in both the bead proteome and PFL appear in the top 2,000 out of 10,623 IPI numbers of the PFL, when proteins are ranked from highest to lowest detection frequency. In contrast, only a small fraction of the bead proteome proteins are found in the bottom (low frequency) end of the PFL (see Figure 29C). This shows that most contaminants identified in the bead proteome are associated with a high frequency in the PFL. In addition, for each sequential PFL “10%” segment, i.e. all proteins associated with a PFL frequency range between 90-100%, 80-90% etc., the number of bead proteome proteins versus the total number of proteins found in that segment were compared. This shows that almost all proteins with a high frequency of detection in the PFL (>60%) are also listed in the bead proteome, while most proteins with a low frequency of detection in the PFL (<20%) are not (see Figure 29D). These data underline the positive correlation between the PFL and the bead proteome and validate the utility of the PFL approach for predicting contaminant proteins. A major advantage of the PFL, as compared with the previous “static” bead proteome, is that it provides an annotation of proteins that is both customisable to reflect the details of individual experiments and updatable. Hence it will increase in accuracy as new data are added to the data repository.

5.3.3 Filtering of Protein Frequency Library using Experimental Parameters

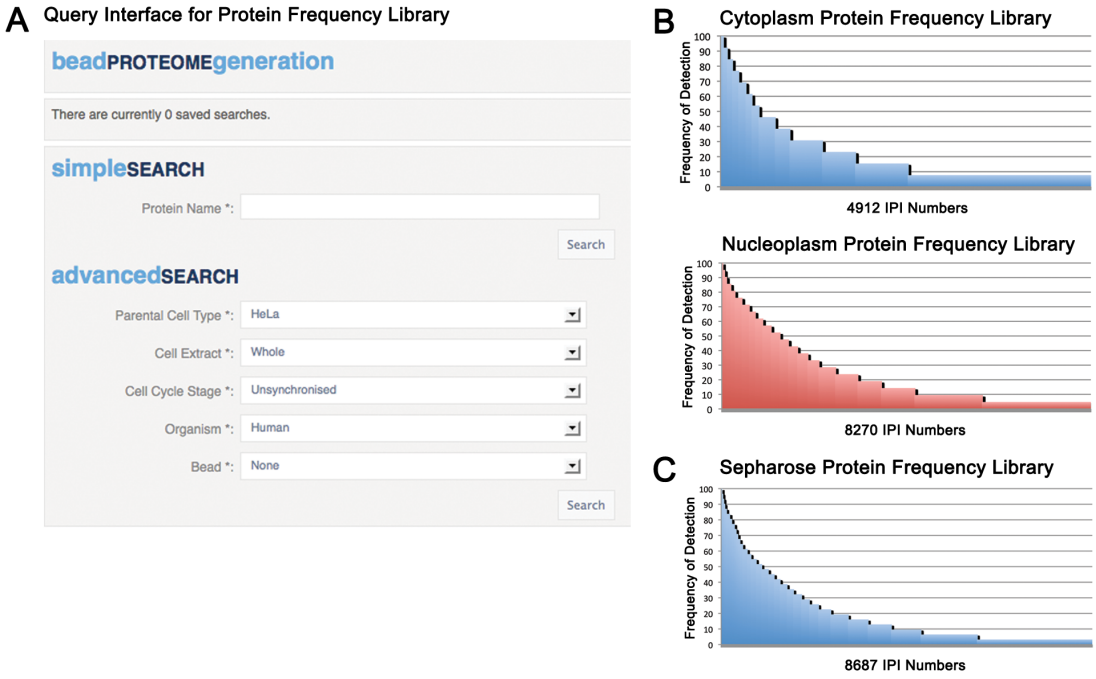


Figure 30: Filtering of PFL using experimental parameters ("dimensions").

(A) Using a web-based interface, any individual dimensions within the data model (corresponding to experimental parameters recorded in the database) can be used in conjunction with the OLAP cube to create a customised PFL. This is illustrated here for the dimensions; (B) Cell extract (cytoplasmic and nuclear) and (C) Affinity matrix (sepharose beads) used for pull-down experiments.

The use of the OLAP cube, and its range of measures and dimensions, provides a dynamic list of contaminants that can be customised for individual experiments. Figure 30A shows an example of an interface in PepTracker that can be used to flexibly specify the parameters (in principle, drawing on all of the dimensions that were incorporated into the cube) on which the library can be filtered so that an analysis is customised to the detailed conditions used for a specific pull-down experiment. Here, the PFL has been filtered using the dimensions "cell extract" (see Figure 30B) and "bead type" (i.e. type of affinity matrix) (see Figure 30C). Thus, among all 38 SILAC pull-down experiments in the data repository, only the ones that were performed with either a specific type of extract (e.g. cytoplasmic or nuclear extract) (see Figure 30B), or with a specific type of bead (e.g. sepharose beads) (see Figure 30C), were used to generate a customised PFL. This customisation feature of the PFL avoids the need to have a large set of control experiments that exhaustively cover every possible

experimental parameter analysed, by combining the different parameters associated with each experiment in the data repository and thus increasing the value of each individual dataset. The PFL is thus applicable also for the analysis of low throughput co-IP experiments, when high-throughput bioinformatics analysis techniques, aiming to discard contaminants, cannot be applied.

5.3.4 Application of PFL to Analysis of Multi-Protein Complexes

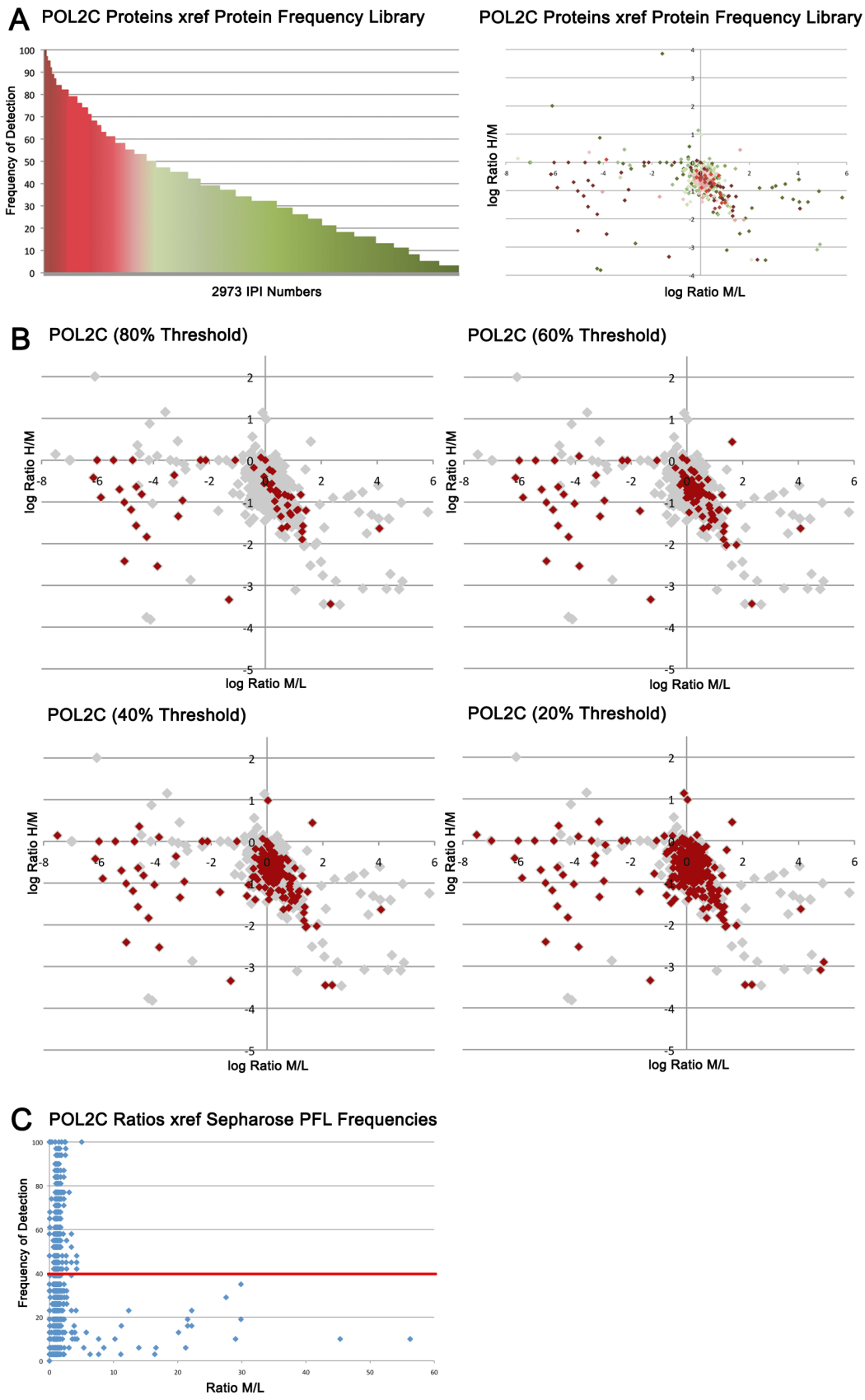


Figure 31: Application of PFL in the identification of specific protein interactors.

(A) Cross-reference between the customized "sepharose" PFL data (as in Figure 30C), and the GFP-Pol2C dataset. Continuous colour coding from red (highest) to green (lowest) is used to depict frequency of protein detection (left panel). On the right panel, the same colour coding is applied to the log_2(H/M) against log_2(M/L) SILAC ratio plot of GFP-Pol2C dataset (plot as shown in Figure 28C). (B) Comparison of arbitrary threshold values (80%, 60%, 40% and 20% detection frequency in PFL), to visualise the most frequently detected proteins in the PFL (highlighted in red) on the log_2 plot of SILAC ratios for the same pull-down experiment shown in (A). Lower threshold values result in highlighting of larger number of proteins. (C) The graph shows the PFL frequency (Y-axis), plotted against the SILAC M/L ratio (X-axis) for each protein group in the Pol2C dataset. A red line is drawn indicating the minimum suitable PFL threshold that includes all protein groups with a high M/L ratio in the likely set of putative interaction partners.

The PFL was applied to analyse the SILAC data from the GFP-Pol2C pull-down experiment (see Figure 31A). As this was performed with sepharose beads, the PFL was filtered to generate a sepharose PFL, as in Figure 30C. A subset of the sepharose PFL library is shown, where only proteins that are identified in the GFP-Pol2C dataset are displayed, i.e. a cross-reference between the Pol2C dataset and the sepharose PFL, which gives a total of 2,973 IPI numbers. A continuous colour coding (from red to green) was applied to the graph, representing proteins with highest detection frequency (red) to the proteins with lowest detection frequency (green) (see Figure 31A, left panel). The same high (red) to low (green) colour coding was then applied to the log_2(H/M) against log_2(M/L) ratio plot (see Figure 31A, right panel). Proteins with high frequency of detection (red) cluster around the origin, while the proteins with lower frequency of detection (green) spread further across the graph. This illustrates the strong positive correlation between proteins that show a high frequency of detection and proteins that cluster around the origin in this IP experiment, which is the expected behaviour of contaminant proteins.

Next, the sepharose PFL was used to isolate within the GFP-Pol2C dataset a group of proteins predicted to include predominantly contaminants. This was done by (i) establishing a threshold value for protein detection frequency and, (ii) highlighting all proteins in the dataset that show a frequency of detection above that threshold. A threshold value of 100% corresponds to only those proteins detected in every dataset in the library. A threshold value of zero instead would include every protein identified in any dataset. Therefore four intermediate frequency thresholds were investigated, corresponding to 80%, 60%, 40% and 20% frequency of detection. Each was applied to

$\log_2$(H/M) versus $\log_2$(M/L) ratio plots of the GFP-Pol2C dataset and compared (see Figure 31B). As the threshold value for the frequency of detection decreases, the number of proteins included in the subset of putative contaminants (highlighted in red on the graphs), increases. The majority of these proteins cluster either at the origin, or on the left quadrants of the graph, exactly as expected if they are indeed contaminants. External contaminants, such as keratins, are always present in the left hand quadrants. At lower threshold values, the probability that some specific interacting proteins are also highlighted is increased. By plotting the PFL frequency value against the M/L SILAC ratio for every protein in the dataset (see Figure 31C), a threshold value of 40% was chosen, because it retains the main stable interaction partners of the bait and selects a suitable subset of clustered contaminants for further normalisation of the GFP-Pol2C dataset, as described below. The choice of an optimal frequency threshold may vary for different experiments. However, the threshold value used is expected to become lower as the number of experiments used to generate the PFL increases. Although it is currently not possible to calculate accurately the minimal number of independent experiments required to provide a reliable PFL, based upon current experience it was estimated that at least 10-15 independent pull-down experiments using different baits constitute a basic requirement.

5.3.5 Use of PFL to Normalise Datasets

Ideally, samples for SILAC analysis are prepared identically, with no variability in experimental conditions and with precisely equal amounts of labelled samples mixed before MS, which should lead to a normal distribution of SILAC ratios centred on exactly zero. However, in practice slight variations in experimental conditions, e.g. pipetting accuracy etc., are unavoidable, resulting in minor variations in SILAC ratios and hence in a ratio distribution whose mean deviates from zero (see Figure 28, B and C). While this generally does not compromise the interpretation of data within a given experiment, it can complicate the accurate comparison of separate datasets, i.e. either biological replicates or independent experiments. Accurate comparison of separate experiments thus requires that datasets are normalised objectively to compensate for intrinsic variations in SILAC ratios.

The MaxQuant software provides a method of data normalisation that is based on the whole dataset in a specific experiment being analysed and this assumes that most

proteins should not change between conditions. However, in a SILAC pull-down experiment it is expected that specific interacting proteins should change between the three conditions: L, M and H. Thus, the PFL is used to normalise datasets by isolating a group of proteins that can confidently be predicted as mostly contaminants and, hence, whose log SILAC ratios should be exactly zero.

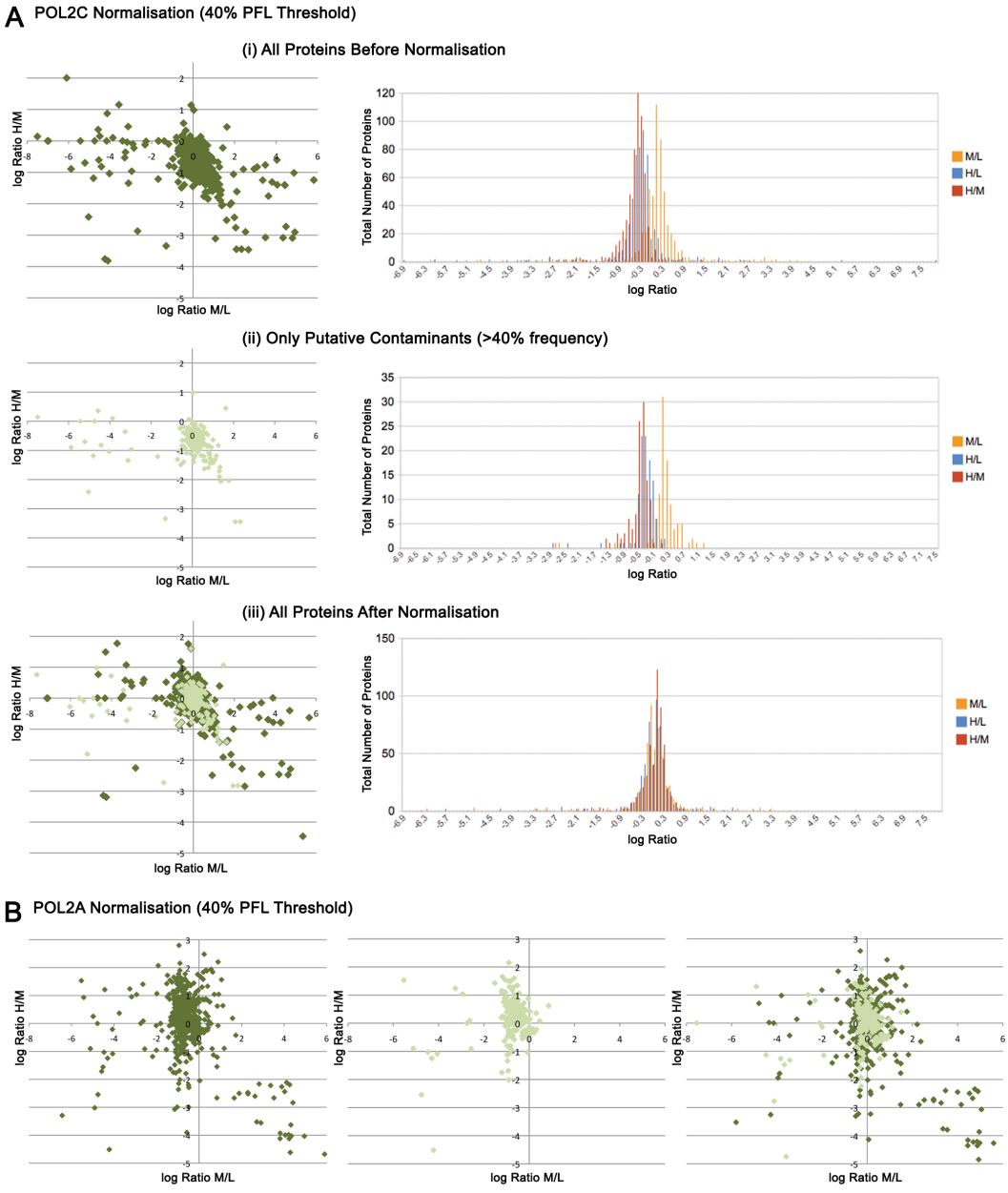


Figure 32: Normalisation of datasets using the PFL.

(A) Graphs show log_2 SILAC ratio plots of total proteins identified from co-IP using GFP-Pol2C as bait (i) before normalisation or threshold analysis (ii) after application of a 40% "sepharose" PFL threshold filter, with the plot now showing only putative contaminants (light green), i.e. proteins with PFL values over 40% (iii) total dataset re-plotted after normalisation to set median SILAC log_2 ratio value of predicted

contaminants to zero. Predicted contaminants are shaded in light green and other proteins shown in dark green. The effect of this normalisation procedure on the Gaussian ratio distribution curves for the three separate M/L, H/L and H/M values recorded in the triple SILAC analysis is shown in parallel on the right for panels (i)- (iii) in the form of SILAC ratio distribution histograms (as in Figure 28B). (B) Repeat of the normalisation procedure shown above using data from a separate triple SILAC co-IP experiment using antibodies to an endogenous protein (Pol2A) rather than a GFP-tagged bait.

The normalisation process is illustrated for the GFP-Pol2C dataset (see Figure 32A). Using the sepharose PFL with a threshold value set to 40% frequency, the resulting proteins with frequency above 40% were isolated within the dataset (see Figure 32A, middle graph), and their median value of SILAC ratios calculated for all three conditions (i.e., M/L, H/L and H/M). MaxQuant non-normalised SILAC ratios were used and the SILAC ratios of external contaminants, e.g. keratins, were excluded from the normalisation process. SILAC ratio values for all proteins in the dataset, including putative contaminants and specific interactors, were then divided by the corresponding median value. This normalises the median log ratio value for the predicted contaminant group to exactly zero (see Figure 32A, right panels). The cluster of contaminants is thereby centred on the origin of the graph (see Figure 32A, bottom graph). This normalisation process does not alter the positions of proteins relative to each other within this experiment, but rather globally affects the ratio values of all the proteins in the dataset. The same normalisation process can be applied to any dataset, including co-IP analysis of an endogenous protein, as shown for the dataset from SILAC affinity-purification of endogenous RNA polymerase II subunit A (Pol2A) (see Figure 32B).

5.3.6 Comparative Analysis of Normalised Datasets

Next the data analysis workflow described above was used to analyse normalised GFP-Pol2C pull-down and endogenous Pol2A co-IP triple SILAC experiments. A customised sepharose PFL combined with a frequency threshold of 40% was used (i) to highlight putative non-specific contaminants and (ii) to normalise the datasets. Figure 33 shows the GFP-Pol2C and endogenous Pol2A datasets plotted as $\log_2$(H/M) against $\log_2$(M/L) ratios after normalisation (see Figure 33 A and B, respectively). Each point represents the normalised median SILAC ratio value for all quantified peptides assigned to that protein. Bait proteins are shown in red. Core subunits of RNA polymerase II are shown in blue. The predicted contaminant-enriched group, i.e. proteins that show a

frequency of detection above 40% in the Sepharose PFL, are shaded in light green, and all other proteins shown in dark green.

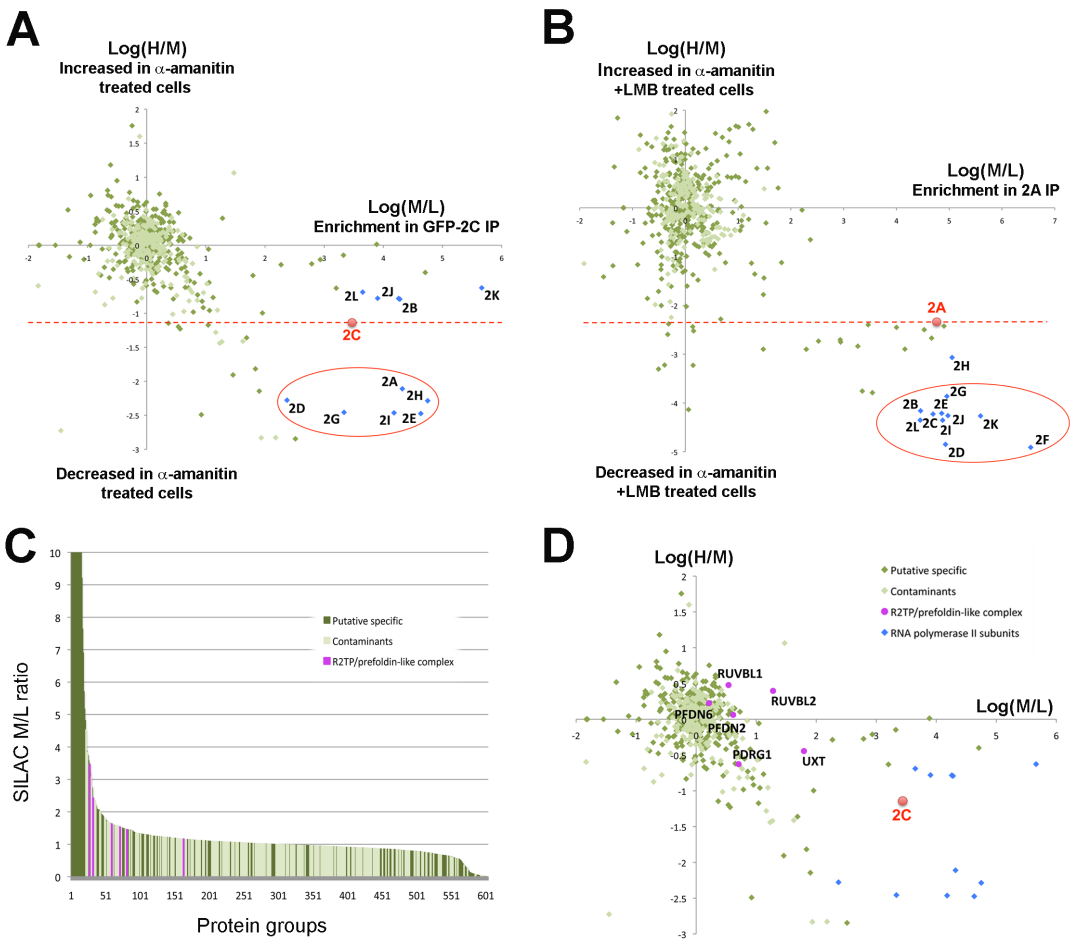


Figure 33: Analysis of protein interaction dynamics using normalised datasets.

Graphs A and B are log_2 SILAC M/L versus H/M ratios comparing normalised datasets from triple SILAC experiments analysing proteins specifically interacting with either (A) GFP-RNA polymerase II subunit C (Pol2C) or (B) endogenous RNA polymerase II subunit A (Pol2A). Each point represents the normalised median SILAC ratio value for all quantified peptides assigned to that protein. Bait proteins are shown in red. Core subunits of RNA polymerase II are shown in blue. A threshold PFL value of 40% was used and all proteins with a 40% or greater frequency value are shown in light green. Dotted red line shows an alternative X-axis defined by the behaviour of the bait protein. Proteins within red ovals are RNA polymerase II subunits whose specific interaction with the bait shows a decrease of 2-fold or more. (C) Identification of specific protein interaction partners with low M/L SILAC ratios using PFL frequencies. The graph shows each protein group identified in the Pol2C dataset plotted on the X-axis and the normalised median SILAC value for that protein group plotted on the Y-axis (similar to Figure 28A). It has been colour coded to highlight all protein groups with a PFL value below 40% in dark green while protein groups showing a frequency value above 40% are shown in light green. Proteins belonging to the R2TP/prefoldin-like complex are highlighted in purple. (D) Same graph as Fig. 8A with the proteins of the R2TP/prefoldin-like complex highlighted in purple.

Gene names	Accession numbers	Nb of unique peptides	Sequence coverage [%]	No. of Pept. Quantif.	SILAC M/L ratio	Log_2(H/M) vs 2C	(H/M) St Dev [%]
POLR2K	IPI00023975.1	3	53.4	3	56.2	0.51	3.4
POLR2L	IPI00003311.1	4	74.6	9	13.9	0.45	8.5
POLR2J	IPI00003310.2;IPI00873238.1;IPI00291359.3;IPI00884938.1;IPI00878433.1;IPI00553186.2;IPI00744926.1;IPI00472231.7;IPI00556199.1;IPI00016841.3;IPI00748167.1;IPI00879159.1;IPI00880135.1	4	57.5	21	21.2	0.35	15.5
POLR2B	IPI00027808.1;IPI00873948.2;IPI00894355.1;IPI00894141.1;IPI00418797.4;IPI00026445.3;IPI00184886.3;IPI00894524.1;IPI00894248.1	81	62.1	272	21.5	0.35	16.6
POLR2C	**IPI00018288.1**	**17**	**85.8**	**73**	**12.4**	**0**	**24.1**
POLR2A	IPI00031627.3;IPI00385524.1;IPI00419565.3;IPI00383337.1;IPI00784155.1	116	61.4	215	22.2	-0.97	36.1
POLR2D	IPI00007283.1	6	62	5	5.7	-1.14	26.4
POLR2H	IPI00003309.4;IPI00791019.1;IPI00790361.1;IPI00791273.1	8	58	12	29.9	-1.15	31.5
POLR2G	IPI00218895.6	11	74.4	14	11.2	-1.32	35.7
POLR2I	IPI00006113.1	7	78.4	9	20.1	-1.33	42.1
POLR2E	IPI00291093.3	10	58.1	23	27.6	-1.34	34.8

Table 5: Comparison of Peptide Data Quality for RNA Polymerase II Subunits.

All known RNA polymerase II subunits (Pol2A–Pol2L) except Pol2F were identified and quantified in the SILAC co-IP using GFP-Pol2C as bait. They all show high sequence coverage and a large number of peptides identified and quantified, underlining the quality of the data. The bait protein, GFP-Pol2C, is bold. Log_2(H/M) ratios of all subunits are normalized in the table so that log_2(H/M) of Pol2C is 0. Subunits are listed above the bait protein when their interaction with the bait was increased upon α-amanitin treatment (log_2(H/M) versus 2C > 0), whereas subunits are listed below when their interaction with the bait protein was decreased upon α-amanitin treatment (log_2(H/M) versus 2C < 0). Interactions are considered significantly affected when log_2(H/M) versus 2C > 1 or log_2(H/M) versus 2C < −1 (equivalent to a change in value of 2-fold or greater).

GFP-Pol2C interaction partners were analysed in U2OS cells either with, or without, α-amanitin treatment (see Figure 33A), while endogenous Pol2A interaction partners were analysed in U2OS cells either with, or without, combined α-amanitin and

leptomycin B (LMB) treatment. In the Pol2C co-IP experiment, eleven out of the twelve known RNA polymerase II subunits (Pol2A-Pol2L), except Pol2F, were detected, with high sequence coverage and a large number of peptides identified and quantified (see Table 5 for GFP-Pol2C dataset), underlining the quality of the data.

If the pull-down efficiency is the same between the two conditions tested (+/- α-amanitin), the $\log_2$(H/M) ratio should be zero for the bait protein. In practice this is often not the case, due for example to variations in expression levels, accessibility and/or fractionation efficiency induced by the treatment. Hence, the bait protein has been used as a reference point to draw a second X-axis such that proteins falling above the new X-axis line indicate increased interaction with the bait and proteins falling below indicate decreased interaction as a result of the treatment. Here, interactions were considered as significantly affected when a two-fold or greater change was observed upon treatment (see Table 5). The GFP-Pol2C dataset shows partial disassembly of the RNA polymerase II complex after α-amanitin treatment (see Figure 33A), because GFP-Pol2C interaction with many RNA polymerase II subunits, including 2A, 2D, 2E, 2G, 2H and 2I, is significantly decreased after α-amanitin treatment (proteins within the red oval, Figure 33A). However, some subunits remain associated, and new protein interaction partners were also identified, suggesting that intermediate sub-complexes are formed upon α-amanitin treatment. The same approach was applied to analyse the Pol2A dataset, showing that Pol2A interaction with all RNA polymerase II subunits, except Pol2H, is decreased after treatment with both α-amanitin and LMB (see Figure 33B, proteins within the red oval). A more detailed analysis and discussion of these data characterising the formation of sub-complexes during RNA polymerase II assembly is presented elsewhere (Boulon et al., 2010b).

Importantly, while high SILAC M/L ratios unambiguously identify specific interaction partners, the application of the PFL to the dataset can help identify additional specific interaction partners otherwise missed because their lower SILAC ratios overlap with non-specific contaminants. This overlap is particularly visible for proteins with a SILAC M/L ratio <3 (see Figure 33C, dark and light green columns). By highlighting all predicted contaminants (frequency of detection > 40%), the PFL approach helps to focus on the remaining putative specific interaction partners. For example, many

proteins of the R2TP/prefoldin-like complex, i.e. UXT, RUVBL1/2, PFDN2/6 and PDRG1, are not identified in the Pol2C dataset with high SILAC M/L ratios but show a frequency value below 40% (see Table 6 and Figure 33D, purple data points). Interestingly, the R2TP/prefoldin-like complex has been connected to the RNA polymerase II complex (11, 36). This shows that these proteins are indeed bona fide interaction partners of Pol2C that would have been overlooked in the analysis without the PFL.

Gene names	Accession numbers	Normalised SILAC M/L ratio	PFL frequency value
KRT19	IPI00479145.2	5.03	100
UXT	IPI00170862.1;IPI00002646.1;IPI00553080.1	3.85	16
ACTN4	IPI00013808.1;IPI00908458.1;IPI00845465.1;IPI00908776.1;IPI00793285.1;IPI00903019.1;IPI00018829.2;IPI00217047.4;IPI00217048.1;IPI00217044.1	3.43	52
RUVBL2	IPI00009104.7;IPI00909925.1	2.70	26
TUBB8	IPI00292496.1	2.01	77
PDRG1	IPI00027887.4	1.81	16
HIST2H2AB	IPI00216730.3;IPI00829588.1	1.77	81
PFDN2	IPI00006052.3	1.7	26
VIM	IPI00418471.6;IPI00552689.1;IPI00465084.6;IPI00793184.1;IPI00013164.4;IPI00910602.1;IPI00021751.5;IPI00217507.5;IPI00869219.1;IPI00853115.1;IPI00908745.1;IPI00237671.9;IPI00868727.1;IPI00166205.2;IPI00477227.3;IPI00001453.2;IPI00853283.1;IPI00744385.2;IPI00909238.1	1.62	97
RUVBL1	IPI00021187.4;IPI00788942.1;IPI00902501.1;IPI00796459.1	1.61	29
FLNC	IPI00178352.5;IPI00413958.4;IPI00455021.3	1.44	77
PFDN6	IPI00005657.1	1.28	13

Table 6: Embedding of Putative Specific Interaction Partners within Contaminants.

A selection of protein groups from the Pol2C dataset with low SILAC M/L ratios (<5) are listed with their PFL frequencies and ranked by M/L ratio from highest to lowest. Protein groups with a PFL value below the threshold (40%), and belonging to the R2TP/prefoldin-like complex, are shaded in grey.

In summary, the PFL approach combined with triple SILAC experiments is shown to provide an effective and flexible workflow for the detection and analysis of specific interactions within multi-protein complexes.

5.3.7 PFL Viewer

To enable researchers from external laboratories to benefit from the PFL concept a web-based interface was created providing access to the PFL generated by the Lamond Laboratory. This interface is particularly important for researchers who do not have the resources or technical expertise to generate their own PFL. Furthermore, researchers who have limited experience in carrying out pull-down experiments or limited access to other researchers experiments may not have the ability to compile enough datasets to generate a reliable PFL to distinguish between genuine interaction partners and contaminants. The Lamond Laboratory benefits from having data generated by multiple researchers over time using a variety of parameters. Hence, the PFL generated by the Lamond Laboratory can be useful in a variety of experiments and is continuously being improved. Allowing external researchers access to the PFL via a web interface would be providing the research community with a useful resource.

The PFL interface was created in conjunction with an honours project student. The honours project student, Laurence Hole (School of Computing, University of Dundee), was supervised over a 9-month period and carried out work that contributed to the development of a GUI for the PFL. The use case for this tool was:

Researcher: Navigates to the website: http://proteinfrequencylibrary.com.

System: Displays a form allowing a researcher to customise the PFL and/or search for a specific protein(s).

Researcher: Enters a set of filtering criteria and/or proteins.

System: Displays the PFL in graph form to the researcher with the selected protein(s) highlighted.

Researcher: Can choose to download the PFL as an excel spreadsheet.

The above use case has been fully implemented in the PFL Viewer tool. The PFL tool is built on top of the multidimensional OLAP cube. Laurence's work centred on creating a library in Python, which allowed the execution of arbitrary Multi Dimensional

Expression (MDX) queries against the OLAP cube. This was accomplished by creating an abstraction layer that used XMLA to transfer data to and from any pre-defined cube. Laurence then focused on creating a basic interface structure to display the data retrieved from the PFL database. This was implemented as a web based prototype. Final work involved converting this prototype to a fully functional tool (http://proteinfrequencylibrary.com) (see Figure 34).

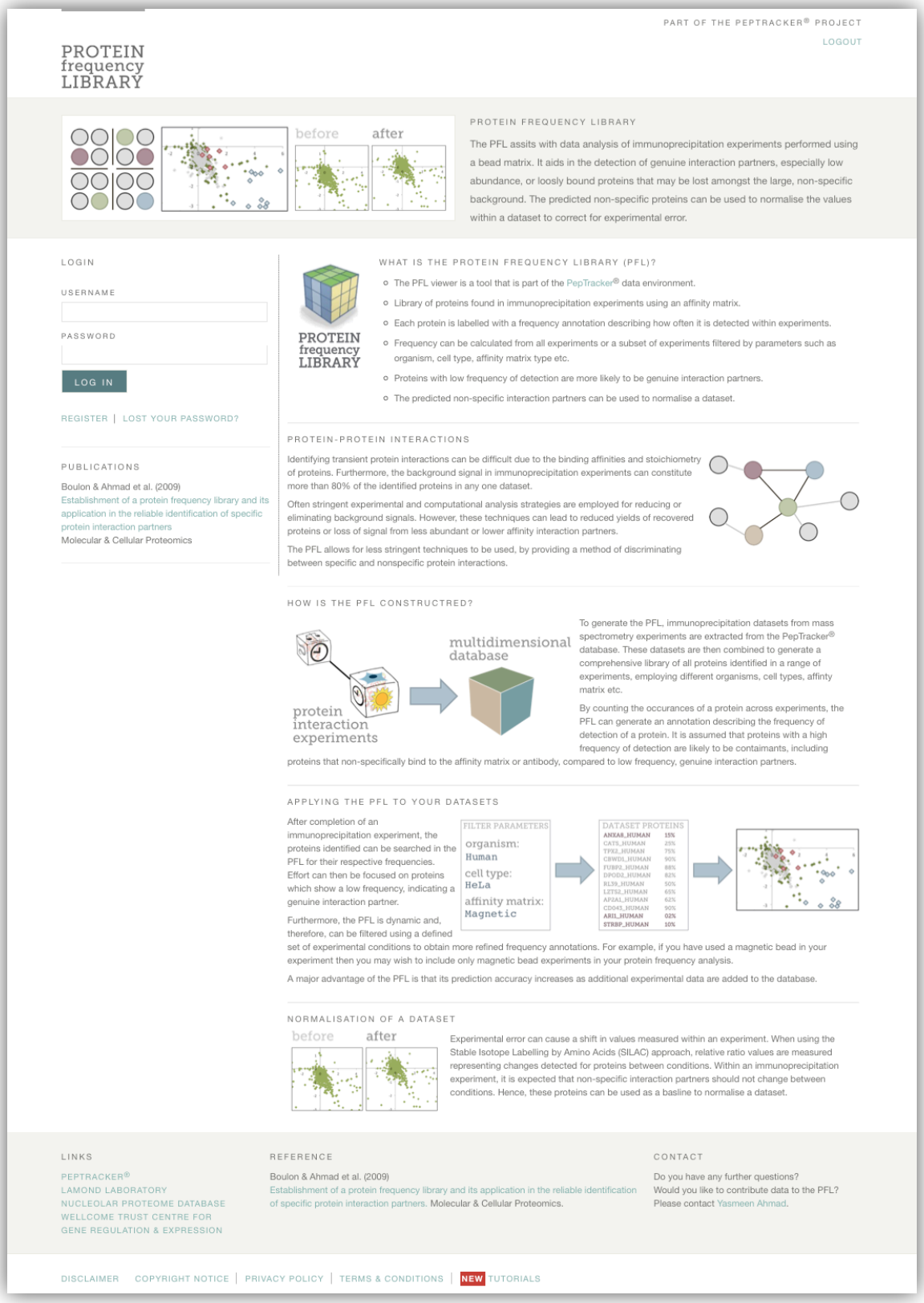


Figure 34: The PFL Viewer tool.

The PFL functionality has also been incorporated into the main PepTracker application. For each dataset, a researcher can choose to apply the PFL to filter proteins from the

display. This has been implemented via a form that has a slider, specifying a selected frequency, and drop down boxes, allowing customization of the PFL (see Figure 35).

Figure 35: PFL filter form within the PepTracker application.

In addition, after having selected to filter a set of proteins based on their PFL frequency annotation, a researcher can choose to normalise their dataset based on these predicted contaminant proteins. As mentioned in 5.3.5 Use of PFL to Normalise Datasets, accurate comparison of separate datasets can be skewed by slight variations in experimental conditions. Hence a method of normalisation has been proposed that is based on a predicted set of contaminants. The PFL functionality in PepTracker automates this normalisation process (see Figure 36).

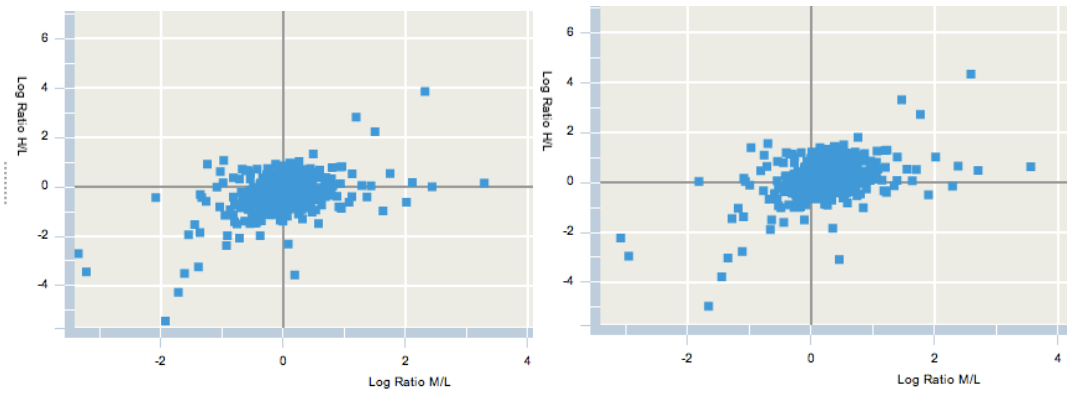


Figure 36: Built-in normalisation functionality.

Original dataset without normalisation is shown on the left. The normalised dataset is shown on the right.

5.4 Technical Implementation

5.4.1 Business Intelligence Application

Due to the complexity of the analysis and the large volumes of data involved, a new approach was required. The alternative approach adopted made use of BI principles, which include methods of leveraging data to provide an informed platform for decision-making (see 1.4.2 Applying Business Intelligence for Super-Experiment Analysis).

The core concept of BI revolves around understanding and modelling data in an appropriate format that makes analysis easier and more intuitive for end-users. BI

technology is designed for rapid interactive response and works particularly well for train-of-thought analysis, whereby response times from queries are rapid enough (one to two seconds) to allow a user to follow a sequence of ideas where each answer can prompt another question. The advantages of rapid response times on productivity have been well understood for many years (Lambert, 1984).

When designing the multidimensional structure, the user model, which is defined by the users' understanding and perception of the data, was translated into a logical model. This logical model is presented in the form of a sun diagram illustrating the relationship between Measures and Dimensions captured in an MS experiment. The measures are numerical values from the experimental data that are of interest to researchers, e.g. ratio, intensity etc. The dimensions define the various groupings (often hierarchical) by which users can aggregate the measures, e.g. treatment, date and cell cycle. The logical model was then represented as a Sun Model (see Figure 37), which shows the measures in the centre of the diagram and dimensions radiating from the centre.

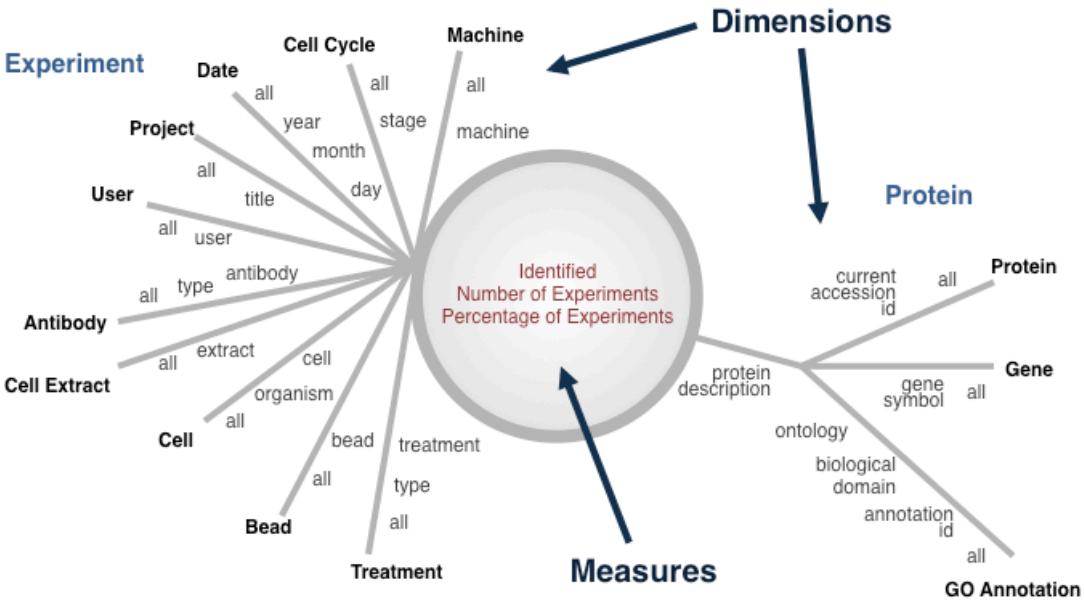


Figure 37: Sun diagram and logical model of SILAC data.

A logical model is presented in the form of a sun diagram illustrating the relationship between Measures and Dimensions captured in a SILAC experiment. The measures are typically numerical values from the experimental data, e.g. "number of peptides". The dimensions define the various groupings (often hierarchical) by which users can aggregate the measures, e.g. cell type, date, cell extract, etc.

The hierarchies in a dimension are symbolised by the levels marked along a dimension line. For example the date dimension is hierarchical and has year, month and day levels. Dimensions, such as “bead type” and “cell extract”, can be used as filters to obtain customised PFLs.

The data for the analysis came from three sources - the relational database within PepTracker (see 4.6.2 Database Development), and local versions of the IPI database, later updated to UniProtKB, and Gene Ontology (GO) (www.geneontology.org) databases. In order to ensure high data quality and consistency of format, the data were extracted from these systems, transformed appropriately and loaded into a central repository (Data Warehouse). During this process, appropriate tables were created to store the data. The measures were incorporated into a Fact table and the dimensions each became a Dimension table. The Fact table maintained a link to all of the related Dimension tables, creating a Star Schema whereby the Dimension tables relate to a central Fact table producing a star shape. This Extract, Transform and Load (ETL) process is characteristic of most BI systems, as is the creation of a Data Warehouse. The data in the Data Warehouse are not updated, but rather appended to when new datasets become available. This method created a Data Warehouse containing historical experimental data that are subject-orientated, non-volatile and well integrated, existing separately from the operational environment of the original data.

In terms of the pull-down datasets, there were a number of decisions that had to be considered carefully to agree how best to transform the data into an accurate representation of the user model. This involved making the determination that proteins should only be included if they were both identified and quantified in a SILAC experiment. In addition, it was decided that proteins should be identified via UniProtKB identifiers, as this is the identifier type common to all datasets searched in the MaxQuant suite (3rd party analysis software) to make protein identifications. Furthermore, due to the continuously updating UniProtKB identifiers, it was decided that proteins should be mapped to the most current identifier. Thus, multiple occurrences of the same protein accession number, in a single experiment, should only be allocated the weighting of a single identification and quantification in the frequency value of any generated protein library annotation. In addition, it was determined that

datasets should be split into separate experiment sets for analysis, where applicable, i.e. single datasets produced from MaxQuant analysis of multiple experiments should be broken down into the constituent experiments for inclusion in the cube.

5.4.2 Multidimensional Database

The next step involved converting the logical model, designed for the PFL, to a physical model, using the SQL Server Analysis Services (SSAS) component of Microsoft SQL Server 2008 R2. This physical model, a multidimensional database, is often known as an OnLine Analytical Processing (OLAP) cube. The OLAP cube is a multidimensional data model that alleviates problems inherent in a relational database by making it easier to select, navigate and explore data. It is also able to provide increased query performance in comparison to a relational database, due to the structure, which supports pre-aggregation of data. Almost all query result times benefit from this type of pre-computation.

The PFL multi-dimensional database is implemented in Microsoft SQL Server 2008 R2. The SQL Server Integration Services (SSIS) component of SQL Server was used to connect and pull in data from the relational database into SQL Server tables. These tables were then used to generate the OLAP cube on the SQL Server using Business Intelligence Development Studio (BIDS). The main PepTracker server communicates with the multi-dimensional OLAP cube using MDX wrapped in XMLA (XML for Analysis) requests. XMLA is an industry standard, XML based Simple Object Access Protocol (SOAP) method of accessing data in analytical systems. An OLAP cube manages data in a cube like structure in which the edges of the cube represent dimensions and the measures are contained within the cube. Data are then extracted from the cube by traversing the edges. Using the hierarchies within the dimensions, users can both drill-down and drill-up to the required level of detail and make use of “slice and dice” operations to change the set of dimensions being viewed. Although an OLAP cube suggests modelling of only three dimensions, in reality data of n-dimensions can be modelled by OLAP in a hypercube structure.

When generating the OLAP cube, the data source, fact tables, dimension tables, relationships between fact and dimension tables and hierarchies had to be specified. Using this information, the cube was then processed, with all of the data aggregated at defined levels within the multidimensional structure. With the use of the OLAP cube

and its modelling of a range of measures and dimensions, it was possible to perform a variety of query and analysis tasks. In order to extract data from the OLAP cube a powerful analytical query language (MDX) is available which allows very complex analytical queries to be expressed with ease (www.microsoft.com/msj/0899/mdx/mdx.aspx). Initially, the PowerPivot component of Excel 2010 (www.office.microsoft.com/excel/) was used to connect to the OLAP cube and extract the required data. Following this a dedicated web interface was created.

5.5 Discussion

The analysis of immunoprecipitation experiments is a challenging task that requires much thought and consideration (ten Have et al., 2011). This chapter has introduced the use of data analysis technology adapted from the field of business intelligence (BI) to improve the reliability of discriminating specific from non-specific protein interaction partners (Boulon et al., 2010a). While this approach is broadly applicable to a wide range of protein interaction analyses, this study focused on describing an enhanced methodology for the analysis of triple SILAC immuno-affinity purification experiments. This identifies genuine protein interaction partners more efficiently and also aids the characterization of changes in protein complexes that can arise either as a result of varied biological conditions, or in response to specific perturbations. To date, there are still relatively few studies that have explored the dynamics of protein-protein interactions using quantitative proteomics-based approaches (Blagoev et al., 2003, Foster et al., 2006). A major aim of the methodology described in this study is to facilitate such analyses. In contrast with other common approaches, this workflow discourages the premature removal of putative contaminant proteins, either experimentally, or in silico. Instead, a comprehensive and inclusive approach has been adopted that takes advantage of the high sensitivity of protein detection now possible using MS-based identification of proteins from model organisms. An interactive analysis is used that integrates several objective criteria to annotate, rather than discard, all proteins in every dataset. This is of particular importance for the detection and characterisation of low affinity and/or low abundance specific protein interaction partners that would otherwise remain undetected amongst the large excess of background contaminants and non-specific interactors.

An important issue in all MS-based protein identification studies is the reliability of protein identification and quantification. While analyses of biological responses are mostly concerned with comparing the differential behaviour of individual proteins, the MS analysis and SILAC procedures directly measure peptides. It must be remembered that the quality of data can differ considerably between separate proteins in the dataset, which can vary in the number of peptides identified and quantified, the total sequence coverage as well as the accuracy and similarity in the SILAC ratios measured for separate peptides assigned to the same protein. Consideration of these parameters can assist with drawing reliable conclusions and they can be incorporated also into the visualisations of the experiments to provide further depth to the analysis of the MS data.

A key feature of the approach described is the generation of a Protein Frequency Library (PFL) that provides a dynamic list of all proteins identified in co-IP experiments and annotation of their frequency of detection. As opposed to the static “bead proteome”, the PFL benefits from continuously being updated with every new experiment that is performed. Thus, addition of new datasets will improve both its reliability and its coverage. The initial PFL described in this chapter contained 10,623 IPI numbers, which corresponds to ~12% of the IPI human proteome. However, this is and can be expanded in the future to cover the entire human proteome, as more datasets from additional co-IP experiments are added, incorporating different conditions and other cell types. In contrast with the previous notion of characterising a set of putative contaminants to eliminate them from the dataset, the PFL approach does not stigmatise any protein as a contaminant. This more accurately reflects the fact that a given protein can interact specifically with certain baits and non-specifically with others. Instead, the PFL provides an objective annotation for all proteins, which predicts their probability of being a contaminant under a defined set of experimental conditions. Applying this annotation to co-IP datasets facilitates discrimination between proteins with high versus low probabilities of being either specific, or non-specific, interaction partners. This is further enhanced by the use of powerful visualisation tools, including the use of colour coding to focus attention on selected sets of proteins identified for further analysis. Furthermore, it provides the ability to flexibly adjust threshold values, as determined by the user, to create optimal settings for each individual experiment.

Another advantage of the PFL approach is that it can be filtered for the parameters from the dataset under analysis, e.g. cell extract, type of affinity matrix etc., to create a customised PFL that more accurately predicts contaminants relevant to each new experiment. The spectrum of parameters available for customisation of the PFL includes all of the dimensions and metadata recorded in the data repository. PepTracker is designed to incorporate a laboratory management tool to facilitate the detailed and consistent recording of metadata from each experiment that can be used directly to generate customised PFLs (see 4.2 MsTrack – Laboratory Information Management System). While the spectrum of dimensions and experimental conditions incorporated in PepTracker is currently focussed on human cells, this can in future be expanded to include a wider range of data, such as other model organisms, and new dimensions, such as detailed genotypes of the cells or organisms being analysed. In addition, the PFL is applicable also to other types of MS analyses not involving SILAC data. For example, it can enhance the analysis of label-free experiments by adding additional objective criteria to identify putative non-specific contaminants.

The generation of the PFL involved adapting advanced techniques from the BI field that deal well with the efficient analysis of large datasets. The core concept of BI revolves around understanding and modelling data in an appropriate format that makes analysis easier and more intuitive for end-users. BI technology is designed for rapid interactive response and works particularly well for train-of-thought analysis, whereby response times from queries are rapid enough (one to two seconds) to allow a user to follow a sequence of ideas where each answer can prompt another question. The advantages of rapid response times on productivity have been well understood for many years (Lambert, 1984). Based on current knowledge, this is the first direct application of such BI technology in cell biology or proteomics research. BI techniques facilitate the analysis of complex data and are essentially discipline agnostic. They have recently been successfully applied, for example, to analyse historical science data, which has enhanced understanding of how Darwin developed the theory of Evolution by natural selection (Kohn et al., 2005). It is hypothesised that wider application of these techniques will be of great utility, not only for proteomics research, but also for other research areas involving the collection and mining of very large datasets, as is now common in biomedical science.

The workflow highlights the need for automation that can deal with the integrated analysis of many large datasets that are inherently multidimensional. The PFL approach can be applied to objectively normalise data and facilitate comparisons of information from separate experiments. The PepTracker environment is capable of storing many consistently annotated datasets and thus presents the opportunity to integrate these datasets, along with associated metadata, to perform what is being termed as a "super-experiment". The PFL represents an example of a super-experiment that incorporates data from a large number of separate immuno-affinity purification experiments. By using this approach to encompass other types of quantitative proteomics experiments it is aimed to expand the super-experiment concept. For example, other types of SILAC and MS analyses provide information about the dynamics of distinct protein properties, such as sub-cellular localisation, turnover and post-translational modifications (Boisvert et al., 2010). Future work will therefore develop the use of BI technology within the PepTracker environment to normalise and mine these combined datasets.

5.6 Distribution of Effort

Lamond Laboratory biologists, primarily Severine Boulon, carried out the experimental bench work during this study. The analysis described in this chapter was a joint effort between Yasmeen Ahmad and Severine Boulon. Yasmeen Ahmad carried out all technical implementation with regards to the BI implementation and supervised a honours project placement student, Laurence Hole, who provided assistance in creating the PFL Viewer.

Chapter 6: Spatial Localisation & Turnover Analyses

6.1 Summary

Measuring the properties of endogenous cell proteins, such as expression level, subcellular localisation and turnover rates, on a whole proteome level remains a major challenge in the post-genome era. Quantitative methods for measuring mRNA expression do not reliably predict corresponding protein levels and provide little or no information on other protein properties. Within the Lamond Laboratory a combined pulse-labelling, spatial proteomics and data analysis strategy was used to characterize the expression, localisation, synthesis, degradation and turnover rates of endogenously expressed, untagged human proteins in different subcellular compartments.

Recent advances in mass spectrometry based proteomics have revolutionized the study of protein dynamics. Mass spectrometry combined with pulsed incorporation of stable isotopes of arginine and lysine has been used to perform quantitative analyses of the rates at which newly synthesized, endogenous proteins appear. The Lamond Laboratory have developed a method using a pulse-labelling strategy combined with SILAC mass spectrometry to characterize the turnover rates and half-lives of proteins in different cellular compartments. HeLa cells were grown in two different SILAC media, containing arginine and lysine, either with the normal 'light' isotopes of carbon, hydrogen and nitrogen (i.e. $^{12}C^{14}N$) (light), or L-arginine-$^{13}C_6{}^{14}N_4$ and L-lysine-$^{2}H_4$ (medium). The medium is then changed for the cells growing in the SILAC medium from "medium" to "heavy", with L-arginine- $^{13}C_6$-$^{15}N_4$ and L-lysine-$^{13}C_6$-$^{15}N_2$, while leaving the cells growing in the "light" medium as a control. Cells were then incubated for 0.5, 4, 7, 11, 27 and 48 hours before being fractionated into cytoplasm, nucleoplasm and nucleoli fractions. Proteins from each fraction and time point were trypsin digested and analysed by LC-MS/MS using an LTQ OrbiTrap. The resulting ratios between light, medium and heavy isotopic forms for each peptide identified were quantified using MaxQuant.

Using this quantitative mass spectrometry and SILAC approach, a total of 80,098 peptides from 8,041 HeLa proteins were quantified, and their spatial distribution between the cytoplasm, nucleus and nucleolus determined and visualised using specialised software tools developed in PepTracker. Using information from ion

intensities and rates of change in isotope ratios, protein abundance levels and protein synthesis, degradation and turnover rates were calculated for the whole cell and for the respective cytoplasmic, nuclear and nucleolar compartments.

The results showed that expression levels of endogenous HeLa proteins varied by up to seven orders of magnitude. The average turnover rate for HeLa proteins was approximately 20 hours. Turnover rate did not correlate with either molecular weight or net charge, but did correlate with abundance, with highly abundant proteins showing longer than average half-lives. Fast turnover proteins had an overall a higher frequency of PEST motifs than slow turnover proteins but no general correlation was observed between amino or carboxy terminal amino acid identities and turnover rates. A subset of proteins were identified that exist in pools with different turnover rates depending on their subcellular localisation. This strongly correlated with subunits of large, multi-protein complexes, suggesting a general mechanism whereby their assembly is controlled in a different subcellular location to their main site of function. A database viewer has been setup to provide access to the data generated during this study through a web-based interface (http://peptracker.com/turnover/).

Chapter 6 describes the analysis of spatial localisation and turnover rates of the HeLa proteome, focusing first on the reasons for protein degradation and turnover, how to experimentally measure protein properties (section 6.2), following with a description of the outcomes of the analysis of the MS data generated and the turnover viewer (section 6.3), implementation of the turnover viewer (section 6.4) and finally a discussion on the importance of protein properties in data analysis (section 6.5).

6.2 Background

Cells can regulate proteins via phosphorylation and other reversible modifications, and through altering protein level by changing the rate of synthesis and/or degradation (Ohsumi, 2006). DNA microarrays are used extensively for analysis of gene expression at the RNA level. Although abundant mRNAs usually result in high protein levels (Lundberg et al., 2010), the general correlation between mRNA levels and protein abundance is often poor (Gygi et al., 1999b). The regulatory complexity of mRNA translation and protein stability emphasises the need for direct measurements of protein levels. Mass spectrometry-based proteomics has emerged as the technology of choice for studying proteins directly, allowing not only identification of proteins and

post-translational modifications, but also quantitative comparisons of how relative protein levels change in cells under different conditions (Walther and Mann, 2010b).

There are two main pathways for intracellular protein degradation, i.e. the proteasome and autophagy-lysosomal systems. The ubiquitin-proteasome pathway identifies proteins for degradation by attachment of poly-ubiquitin tags, which targets the modified proteins for degradation by the proteasome (Clague and Urbe, 2010). In the autophagy-lysosomal system proteins destined for degradation are captured within membrane bound organelles (phagosomes) for bulk digestion (Ohsumi, 2006). Cell growth requires a net increase in total protein and thus higher levels of translation than degradation. Maintaining protein levels at steady state also involves continuous protein synthesis, balanced with degradation. Protein turnover rates can range from under ten minutes to over a hundred hours (Ohsumi, 2006).

Some biological processes involve constant cycles of protein production and rapid degradation. For example, despite continuous synthesis of the tumour suppressor p53, its constant rapid degradation results in low steady state levels under normal cell growth conditions (Lane and Levine, 2010). Upon oncogene activation, degradation of p53 is prevented through sequestration of the E3 ligase mdm2, causing a rapid increase in p53 levels independent of transcriptional activation. Control of protein degradation thus provides a flexible mechanism for the rapid activation of gene expression in mammalian cells.

The turnover rates of specific proteins can vary between different subcellular compartments. Using a combination of pulsed SILAC and fluorescence microscopy, it was shown that HeLa cells constantly import and degrade high levels of free ribosomal proteins in the nucleus. Ribosomal protein stability is dramatically increased upon assembly into ribosome subunits and export to the cytoplasm (Lam et al., 2007). This constant degradation of free ribosomal proteins in the nucleus may allow cells to rapidly upregulate the rate of ribosome subunit production when cell growth rate increases while preventing the accumulation of a large pool of unbound ribosomal proteins. Importantly, this analysis of ribosomal protein turnover shows that protein half-life values based only on analyses of whole cell extracts provide average values that can mask the existence of pools of protein with different properties.

Early studies of protein turnover relied on detecting incorporation of radiolabeled amino acids into newly translated proteins and either analysed bulk cellular protein turnover, or else turnover of individual proteins (Garlick and Millward, 1972). Typically, proteins were labelled with [35S] methionine and pulse-chase experiments used to determine their rate of degradation after blocking protein synthesis, using inhibitors such as cycloheximide. The use of protein synthesis inhibitors raises concerns whether the normal degradation processes, or other aspects of cellular activity, may also be disrupted. Mass spectrometry-based proteomics now allows determination of the turnover rates of large numbers of proteins in single experiments using pulse labelling with amino acids incorporating stable isotopes (Pratt et al., 2002, Doherty et al., 2005, Milner et al., 2006, Schwanhausser et al., 2009).

SILAC (see 1.1.3 Stable Isotope Labelling using Amino Acids in Cell Culture) has been successful for quantitative analysis of cell and organelle proteomes and for comparative studies of protein modifications, and interactions (Walther and Mann, 2010b). SILAC has been used in combination with cell fractionation to generate 'isotope-encoded' subcellular compartments allowing subcellular protein localisation to be evaluated on a system-wide level (Boisvert et al., 2010). This spatial proteomics approach provides a high-throughput assay for the unbiased analysis of changes in subcellular protein localisation arising in response to perturbations such as DNA damage and for comparing protein localisation and responses in cell lines with different genotypes (Boisvert et al., 2010).

Here an enhanced pulse SILAC approach is combined with spatial proteomics to perform a system-wide analysis of protein turnover in cultured human cells. Protein abundance and the rates of protein synthesis, degradation and turnover have been measured in parallel for whole cells and for separate cytoplasmic, nuclear and nucleolar compartments, providing a cell-based functional annotation of the human proteome.

6.2.1 Experimental Design

The experiment was designed and carried out by researchers in the Lamond Laboratory. HeLa cells were grown in parallel in media containing arginine and lysine, either with the normal 'light' isotopes of carbon, hydrogen and nitrogen (i.e. $^{12}C^{14}N$) (light – 'L'), or else with L-arginine-$^{13}C_6{}^{14}N_4$ and L-lysine-$^{2}H_4$ (medium – 'M') for at least

5 cell divisions, resulting in >99% incorporation of the M amino acids in cell proteins (see Figure 38A). The culture media of the cells growing with the M amino acids is then replaced with media containing L-arginine- $^{13}C_6$-$^{15}N_4$ and L-lysine-$^{13}C_6$-$^{15}N_2$ (heavy –'H'). Thus, H amino acids are pulsed into cells with M-labelled proteins for varying times, from 30 minutes to 48 hours. For each peptide at each time point the fraction of H amino acids incorporated, replacing the pre-existing M amino acids, is determined by MS.

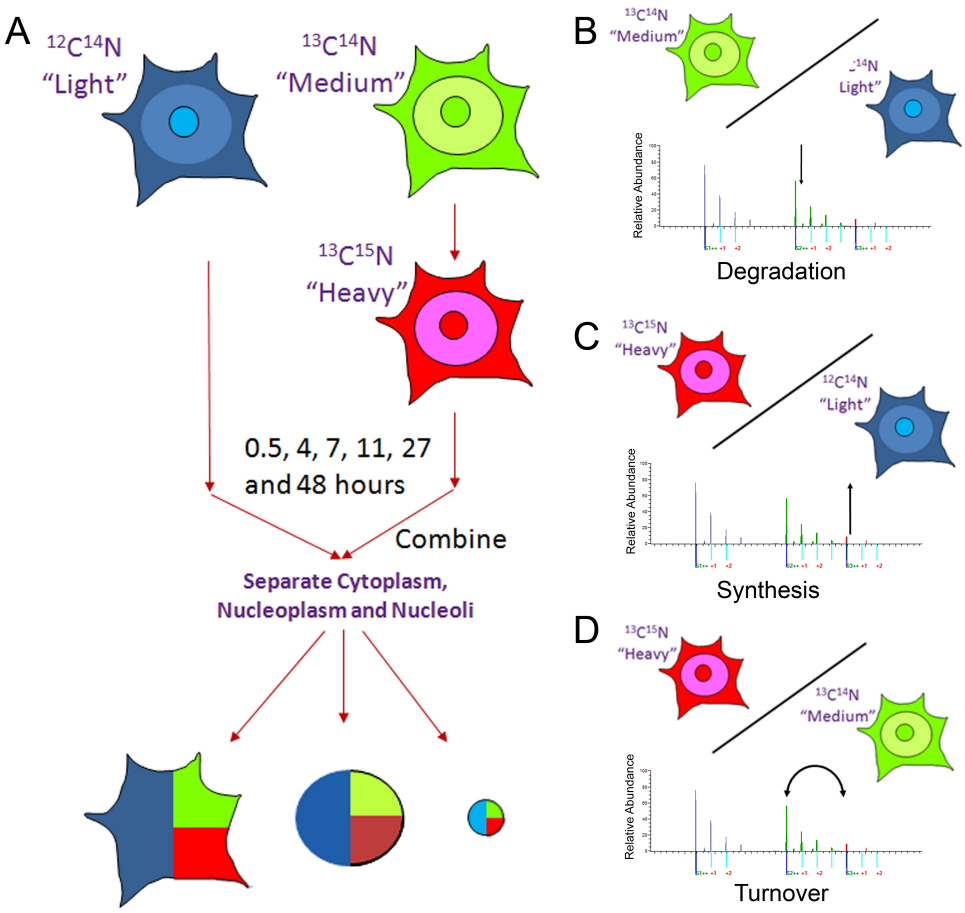


Figure 38: Pulse SILAC method.

A) HeLa cells are cultured in different SILAC media containing either "light" (L), or "medium" (M) arginines and lysines until full incorporation of the amino acids. The medium of the cells growing with the "medium" amino acids is then changed for a "heavy" (H) medium. Cells are then harvested at different times, along with the equivalent cells growing in the "light" medium. Equal amounts of cells are then combined and separate cytoplasmic, nuclear and nucleolar fractions were isolated from each time point. The resulting ratios M/L isotopes over time measures the rate of protein degradation B), increase in the ratio of H/L measures new protein synthesis C) and the change in the H/M ratio measures the rate of net protein turnover D).

Cells were harvested at 0.5, 4, 7, 11, 27 and 48 hour time points following the H amino acid-pulse. A key feature of this pulse-labelling strategy involves, at each time point,

mixing the pulsed cell sample with an equal number of HeLa cells grown in normal (i.e. light – 'L') culture media. This provides an internal control, allows separate measurement of protein synthesis, degradation and turnover rates and facilitates normalisation of the isotope incorporation data, thereby improving the accuracy of the measurements. Moreover, this light sample enables the use of peptide ion intensity to estimate protein abundance, both in the whole cell, and in each subcellular compartment. The decreasing ratio of M/L isotopes over time measures the rate of protein degradation (see Figure 38B), while the increasing ratio of H/L measures protein translation (see Figure 38C) and the change in the H/M ratio measures the rate of net protein turnover (see Figure 38D). The turnover time for each protein was also determined separately by analysis of the crossover between the respective synthesis and degradation curves and these values compared with the turnover values obtained by measuring rates of change in H/M ratio.

The mixture of 50% L cells with 50% H/M cells was fractionated to generate separate cytoplasmic, nuclear and nucleolar fractions for each time point (see Figure 38A). External protein contaminants, such as keratins, will only appear in the L samples because the heavy isotopes used occur at very low levels in the environment. All resulting samples were solubilised with loading buffer, proteins separated using SDS-PAGE and the resulting gels cut into 16 equal pieces, trypsin digested and analysed by LC-MS/MS. Every sample was analysed twice by mass spectrometry and the resulting ratios between light, medium and heavy isotopic forms for each peptide identified were quantified using MaxQuant (Cox and Mann, 2008).

6.3 Results

6.3.1 Protein Identification, Abundance and Subcellular Localisation

Time-resolved protein abundances (hereafter protein profiles) were measured based on the constituent peptide signal intensities of the light sample at each time point (Carrillo et al., 2010). This initial data processing was carried out using the Chromoprot software package.

This analysis has identified and quantitated 80,098 peptides, mapped onto 8,041 endogenous HeLa cell proteins, yielding an average of ~10 peptides per protein. The abundance of each protein was estimated based on the averaged peptide ion

intensities from the control, light sample at each time point (Carrillo et al., 2010). Peptide intensity profiles were normalised from the top three peptides, based on their mean profile intensity.

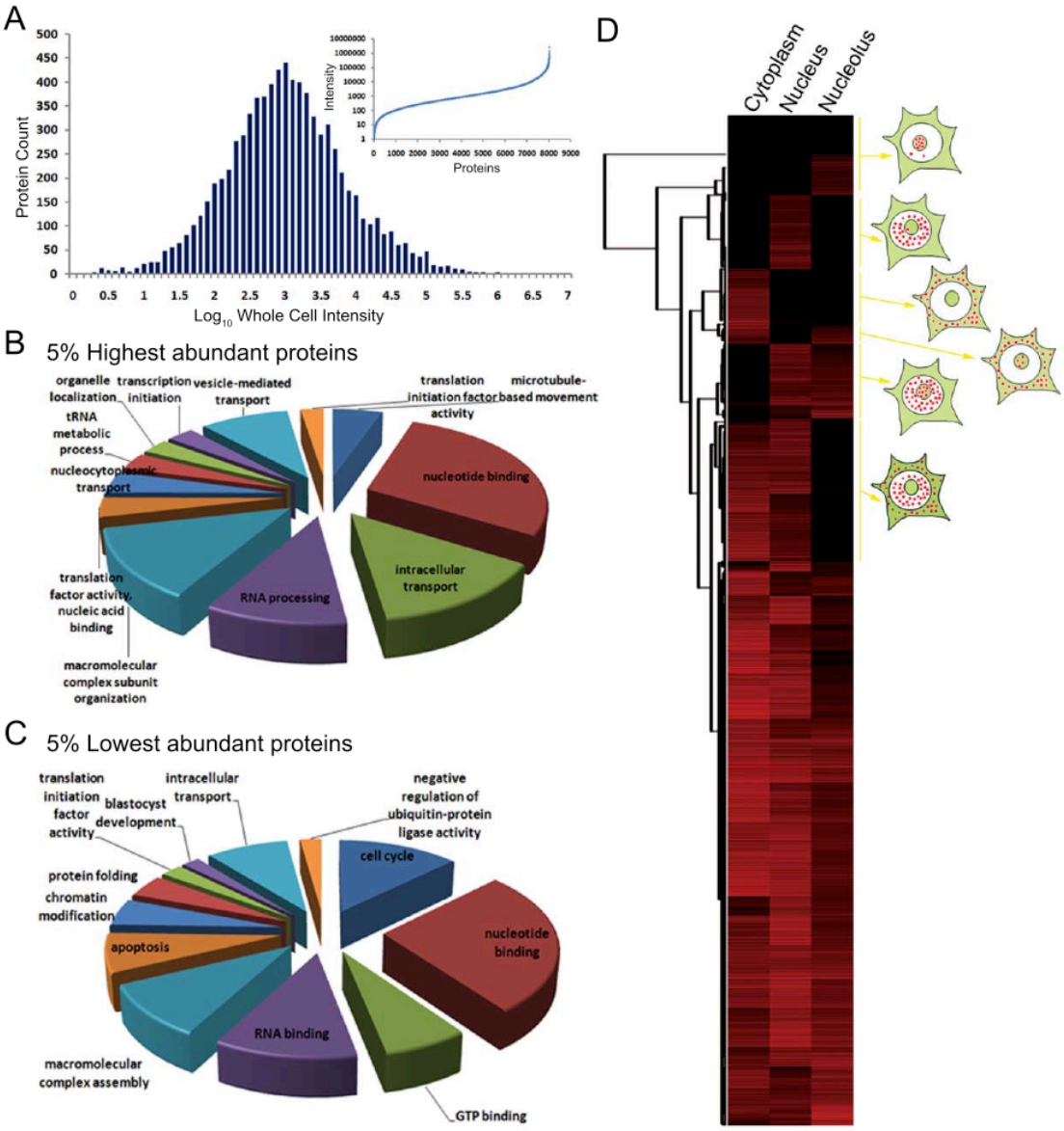


Figure 39: Protein identification, abundance and subcellular localisation.

Peptide intensity profiles normalized from the top three peptides based on their mean profile intensity were used to measure protein abundance. A) A distribution plot with the protein count on the y-axis and bins of 0.1 of the log_{10} intensity values on the x-axis. The inset shows the distribution from the lowest intensity to the highest intensity protein with the intensity on the y-axis and the protein number on the x-axis. B) A gene ontology annotation analysis of the 5% most abundant proteins identified using functional clustering of biological processes and molecular functions (GO_BP and GO_MF). C) A gene ontology annotation analysis of the 5% lowest abundant proteins identified using functional clustering of biological processes and molecular functions. D) A hierarchical clustering was performed using the log_{10} value for intensity for the

cytoplasm, the nucleus and the nucleolus and represented as a heat map. In each case high values are shown in red and low ratios in black.

The protein abundance data span a dynamic range intensity of ~$1x10^7$ following a normal distribution with a mean intensity of ~7,000 (see Figure 39A). This shows the very large variation in copy number of endogenous human proteins expressed from different genes. Known abundant proteins, including nucleophosmin, histones, ribosomal proteins, actin, tubulin, GAPDH and heat shock proteins, were amongst the top 1% highest intensity proteins. Histones are predominantly stably incorporated into nucleosomes with on average ~150 million nucleosomes per human cell. Therefore, as histones showed ion intensities ~1,000,000, it is estimated that proteins with the lowest intensity values have a copy number ~50-150 molecules per cell while the bulk of HeLa proteins are expressed at ~1,000-10,000 copies per cell (see Figure 39A). However, as these estimates are derived from averaging values over the cell population, there could be significant variation in the levels of proteins present at the single cell level.

Gene ontology annotation analysis of the 5% most abundant proteins identified factors involved in nucleotide binding, intracellular transport, RNA processing and macromolecular complex subunit organization (see Figure 39B). Analysis of the 5% lowest abundance proteins revealed functions related to nucleotide binding, GTP binding, RNA binding and cell cycle regulation (see Figure 39C). While both the highest and lowest abundance protein groups had “nucleotide binding” as the largest class, the types of nucleotide binding proteins were different in each case. Thus, many transcription factors were included amongst the very low abundance proteins while histones and hnRNPs were prominent amongst the high abundance proteins. Over 40% of the lowest abundance proteins are either uncharacterized open reading frames (ORFs), or else proteins named only based on their molecular weight, or on a recognizable domain. In contrast, less than 1% of the most highly abundant proteins are uncharacterized ORFs or of unknown function. Overall, these results show that the intensity values measured reflect the relative abundance of the over 8,000 HeLa proteins identified.

The ‘light’ peptide ion intensities measured in the fractionated subcellular compartments allow separate estimation of protein abundance in the cytoplasm,

nucleus and nucleolus, providing a quantitative map of protein localisation within the cell. The absence of ion intensity values in a specific compartment is interpreted as meaning that the protein was present in very low abundance in this location and thus assigns it an intensity value of 0. A hierarchical clustering was performed and visualised as a heat map, using $\log_{10}$ intensity value for each compartment (see Figure 39D). For more than half of the proteins, intensity values were detected in more than one compartment. Relatively few proteins show equal distributions between two or three compartments (see Figure 39D). This suggests that, at steady state, most HeLa proteins are predominantly partitioned into specific subcellular locations. However, this does not exclude that proteins can shuttle between their major site of accumulation and other compartments.

6.3.2 Determination of Protein Turnover

Two methods have been used to evaluate the time point at which 50% turnover has occurred for each protein. The first method, relying on changes in the H/M isotope ratio, directly measures when 50% of the intensity signal for a peptide is M and 50% H, isotope. The corresponding protein turnover is the mean time for 50% incorporation of H amino acids for all peptides identified from that protein. The second method relies upon measuring the separate curves of synthesis and degradation rates for each protein, based on rates of change in H/L and M/L isotope ratios, and then identifying the point at which these curves cross, which corresponds to the time it takes for 50% of the protein to turn over.

An interesting observation from the crossover method is a measured offset value (B') of ~20%. This B' value reflects the fraction of M amino acids remaining in proteins once a steady state level of H amino acid incorporation is established. If H amino acid incorporation completely replaced the pre-existing M amino acids then the B' value should be zero. The fact that it remains at ~20% suggests recycling of M isotope-containing amino acids into proteins. Most likely this results from degradation of the pre-existing pool of exclusively M isotope-labelled proteins within cells at the start of the pulse. Indeed, previous work has reported amino acid recycling from degraded proteins (Davies and Humphrey, 1978). To test whether the amino acid pool settles at approximately 80% H amino acids, the mass isotopomer distribution of peptides with missed trypsin cleavage was analysed to determine the level of peptides that

contained both M + H amino acids. This showed that ~10-20% of peptides with missed cleavages consistently had both M and H amino acids in the same peptide, consistent with a precursor pool of ~80-90% H amino acids. Moreover, it was found for these same missed cleavage peptides that there was virtually no combined M + H amino acids in the same peptide in an experiment where the SILAC medium was changed every hour over the time course. It was infered that without more than one replacement of the cell growth medium during the course of the experiment, the internal amino acid pool is likely not fully replaced with the externally supplied H amino acids.

A simple mathematical model of protein synthesis and degradation developed here demonstrates that recycling of degraded proteins can lead to a non-zero offset in degradation curves. To test this hypothesis, the offset value B was analysed to determine whether it would decrease towards zero if during the time course of the pulse the external media containing H amino acids was repeatedly replaced. This showed that replacing the media containing the H amino acids several times during the time course of the pulse resulted in the offset B reducing to ~ 0, as expected for complete replacement of M with H amino acids. It can be concluded that the intracellular pool of M amino acids either is not fully depleted when the medium is initially replaced, or else is replenished through recycling of amino acids from degradation of pre-existing M-labelled proteins, or both.

Another parameter that was determined was the respective protein half-life values, which represents here the time taken for 50% of the pool of each pre-existing protein species to be degraded. It was noted that this study does not provide half-life values at the single molecule level but rather reflects average values for populations of protein molecules. The half-live values should reflect rates of protein degradation. To take into account the inevitable dilution effect on pools of pre-existing M-labelled proteins as a result of cell growth and new protein synthesis, rather than degradation, the half-lives were calculated using a formula that incorporates the growth rate measured here for HeLa cells growing in SILAC medium. A comparison of the separate protein turnover and half-life values determined in this study showed that they are closely correlated (Pearson Correlation Coefficient 0.54). As the 50% turnover values more directly reflect both protein synthesis and degradation rates, and can be measured more

accurately, the subsequent analyses has been focused specifically on a comparison of turnover values with other protein properties.

6.3.3 Distribution of Protein Turnover

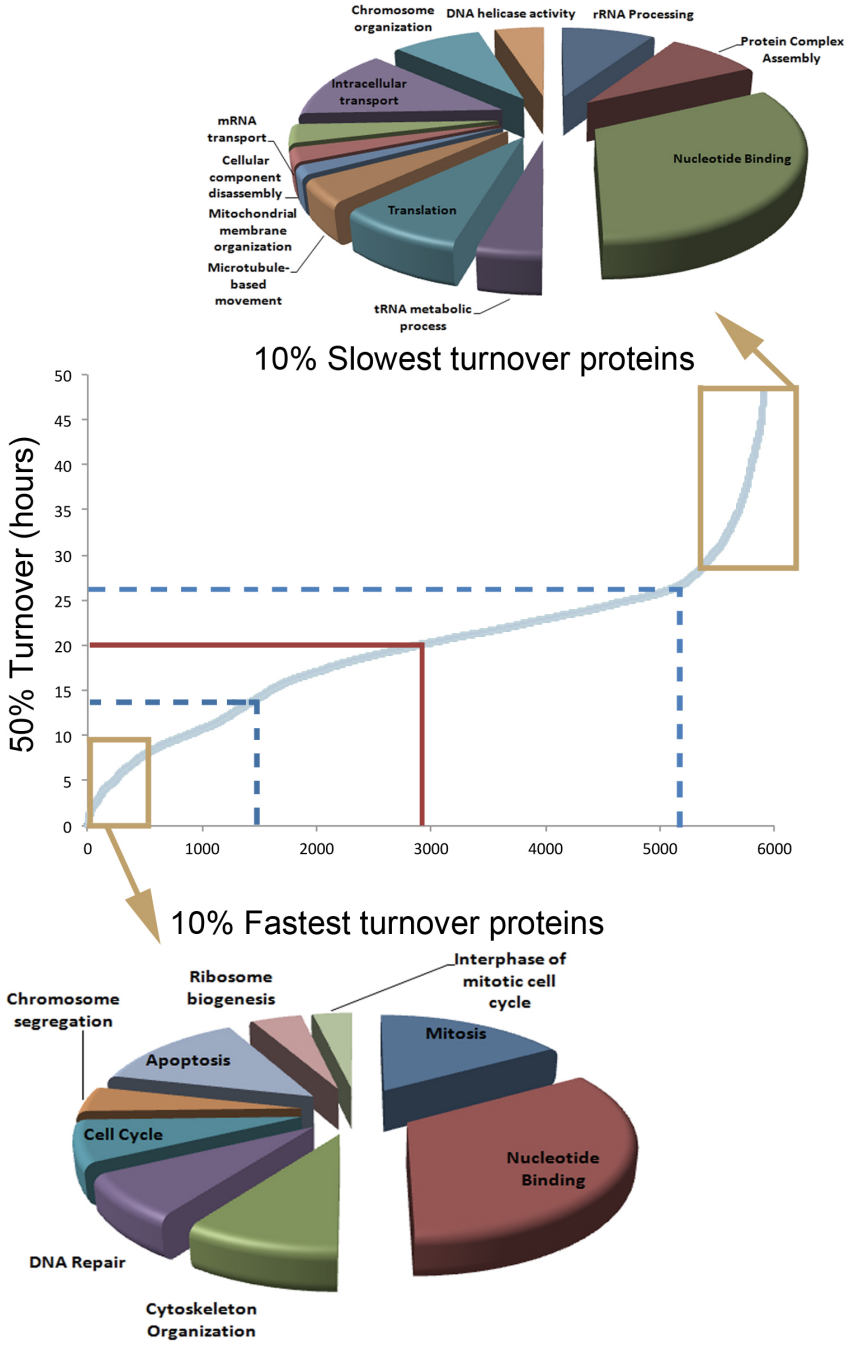


Figure 40: Distribution of protein turnover.

Proteins were sorted on the x axis from fastest to slowest turnover and represented as a scatter plot with the 50% protein turnover value on the y axis. Approximately 60% (blue lines) of the HeLa proteins show a 50% turnover rate within 5 hours of the average of ~ 20 hours (red lines). Functional annotation clustering of gene ontology

terms for the 10% proteins with the fastest (bottom) and slowest (top) turnover rates are shown as pie charts, using the number of proteins as weight for each annotation.

Proteins are sorted on the x-axis from fastest to slowest turnover rate, represented as a scatter plot with the protein turnover on the y-axis (see Figure 40). Approximately 60% of HeLa proteins have turnover values clustered within 5 hours of the average turnover rate of ~ 20 hours (see Figure 40, blue lines). This is close to the cell doubling time under the growth conditions used, consistent with approximate doubling of the protein content as the cell divides (see Figure 40), red line). It takes ~24 hours for 50% turnover of the total HeLa proteome, however, a subset of abundant proteins, including ribosomal proteins, cytoskeletal proteins and histones, have half-lives longer than the mean of ~20 hours.

Functional annotation clustering of gene ontology terms for the fastest and slowest turnover rates showed specific enrichments of proteins with similar functions or characteristics (Dennis et al., 2003, Huang da et al., 2009) (see Figure 40). The slowest turnover proteins have a wide range of functions. However, most are either present in large, abundant and stable protein complexes, such as ribosome and spliceosome subunits, RNA polymerase II, the nuclear pore, the exosome and the proteasome, or else are mitochondrial (see Figure 40, top). In contrast, many proteins with a faster than average turnover are involved in either mitosis, or other aspects of cell cycle regulation (see Figure 40, bottom). This includes protein components of the centromere, proteins with microtubule motor activity, proteins involved in cytoskeleton reorganization and proteins involved in chromatin assembly and condensation. It is noted that this study analysed unsynchronised HeLa cells where only a minor fraction of the cells at any time point would be in mitosis.

6.3.4 Protein Turnover in Different Subcellular Compartments

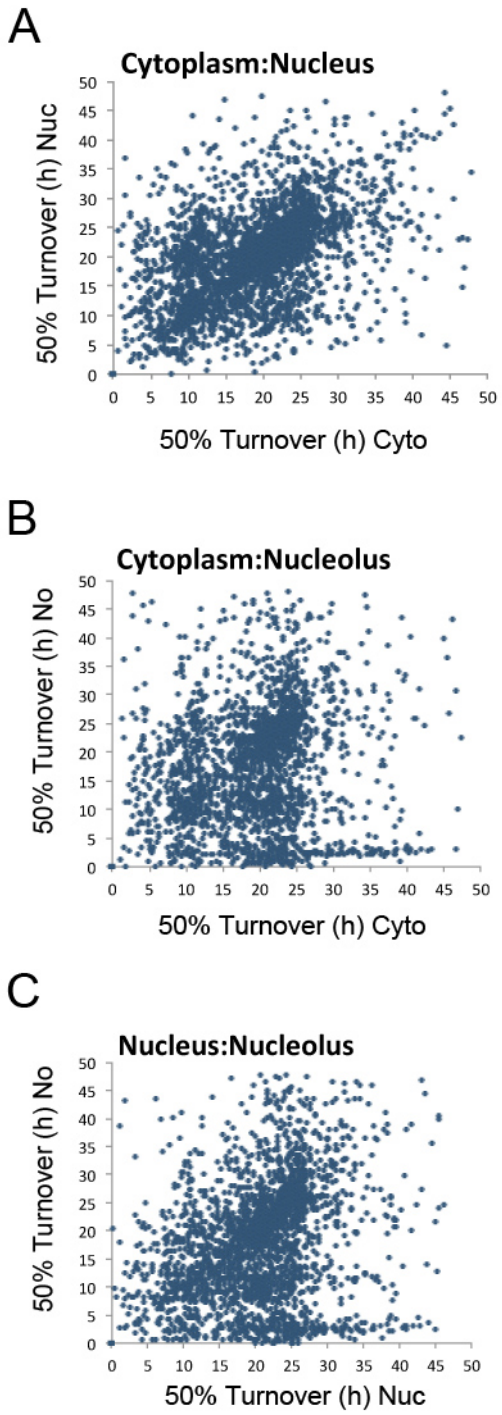


Figure 41: Protein turnover in subcellular compartments.

The turnover data for subcellular compartments are plotted against each other to compare the 50% turnover values for each protein in the cytoplasm versus the nucleus (A), the cytoplasm versus the nucleolus (B) and the nucleus versus the nucleolus (C).

The spatial proteomics approach (Boisvert et al., 2010, Boisvert and Lamond, 2010) was combined with pulsed SILAC to measure the turnover of proteins in the separate cytoplasmic, nucleoplasmic and nucleolar fractions. The turnover data for subcellular

compartments are plotted against each other for comparison (see Figure 41). This shows that most proteins have a similar turnover rate in each compartment, particularly comparing nucleus and cytoplasm (see Figure 41 A versus B and C). Performing correlation analyses between the different compartment shows that the Pearson correlation coefficient between the cytoplasm and the nucleus is 0.67, compared to 0.42 between the cytoplasm and the nucleolus and 0.50 between the nucleus and the nucleolus.

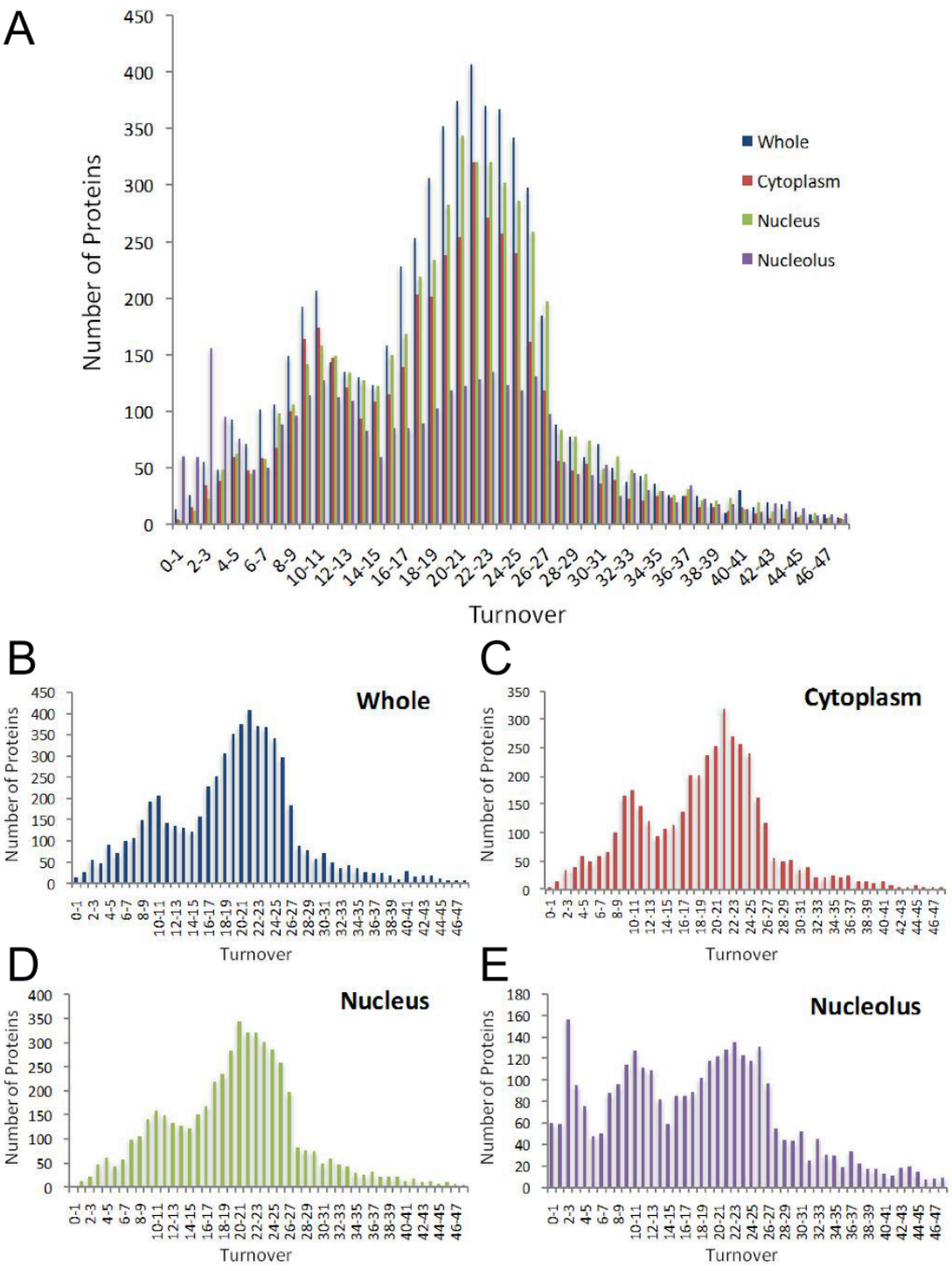


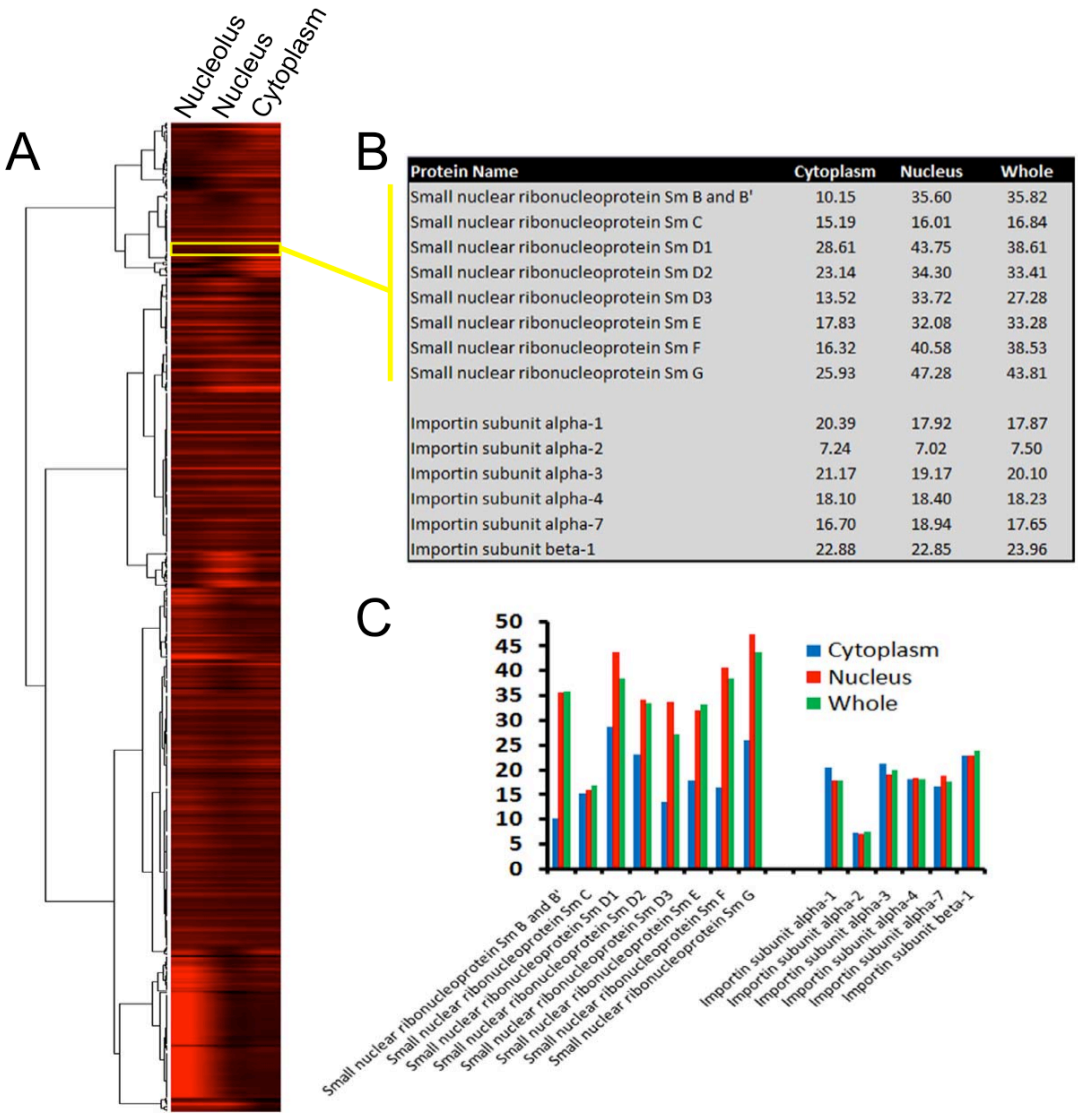
Figure 42: Distribution of protein turnover in subcellular compartments.

A distribution plot with the protein count on the y-axis and bins of 1 hour 50% turnover values on the x-axis for the whole cell (B), the cytoplasmic (C), nuclear (D) and nucleolar (E) proteins, as well as an overlay of all four (A).

The turnover data for subcellular compartments are shown sorted on the x-axis in the same order in each case, reflecting highest to slowest turnover rate when the whole cell protein turnover measurements are compared (see Figure 42A). Again, this shows

that most proteins have a similar turnover rate in each compartment, particularly comparing nucleus and cytoplasm (see Figure 42C versus E), likely reflecting the large amount of nucleo-cytoplasmic shuttling. However, a subset of proteins show differences in turnover rate between the subcellular compartments (see Figure 42A).

Protein turnover follows an apparent bimodal distribution, with a major peak in the number of proteins with a 50% turnover value of ~20 hours, and a minor peak of ~10 hours (see Figure 42C, D & E). The similar distribution of protein turnover rates for the cytoplasm and the nucleoplasm (see Figure 42, C & D), contrasts with the nucleolus, where there appears to be a third peak with a faster 50% turnover rate (<6 hours), while the major peak is centred ~22-23 hours, slower than the whole cell mean turnover value of ~20 hours (see Figure 42A). The nucleolar proteins with the fastest turnover rate are predominantly ribosomal proteins, as previously observed (Lam et al., 2007).



Protein Name	Cytoplasm	Nucleus	Whole
Small nuclear ribonucleoprotein Sm B and B'	10.15	35.60	35.82
Small nuclear ribonucleoprotein Sm C	15.19	16.01	16.84
Small nuclear ribonucleoprotein Sm D1	28.61	43.75	38.61
Small nuclear ribonucleoprotein Sm D2	23.14	34.30	33.41
Small nuclear ribonucleoprotein Sm D3	13.52	33.72	27.28
Small nuclear ribonucleoprotein Sm E	17.83	32.08	33.28
Small nuclear ribonucleoprotein Sm F	16.32	40.58	38.53
Small nuclear ribonucleoprotein Sm G	25.93	47.28	43.81
Importin subunit alpha-1	20.39	17.92	17.87
Importin subunit alpha-2	7.24	7.02	7.50
Importin subunit alpha-3	21.17	19.17	20.10
Importin subunit alpha-4	18.10	18.40	18.23
Importin subunit alpha-7	16.70	18.94	17.65
Importin subunit beta-1	22.88	22.85	23.96

Figure 43: Subcellular clustering analysis of protein turnover.

A) A hierarchical clustering using the 50% turnover values for proteins in the cytoplasm, the nucleus and the nucleolus is shown represented as a heat map. Fast turnover values are represented in red and slow turnover in black. B) A table showing the 50% turnover of the Sm proteins, i.e., subunits of the small nuclear ribonucleoprotein (snRNP) spliceosome and the Importin transport receptor proteins in the three subcellular compartments. C) Graphical representation of the 50% turnover value of each protein in the cytoplasm (blue), the nucleus (red) or the average for the whole cell (green), with the turnover on the y-axis.

A clustering analysis grouped proteins with similar turnover rates in either the cytoplasm, nucleoplasm or nucleoli, represented as a heat map with the protein clusters on the y axis and the subcellular compartments on the x-axis (see Figure 43A). Most proteins showed similar turnover rates in each compartment. Ribosomal proteins provide a clear example of a protein cluster with differing turnover rates between compartments, i.e. fast turnover in the nucleolus (~6 hours), but slow turnover in the

cytoplasm (~30 hours), (see Figure 43A, bottom cluster). Other examples were identified where multiple subunits of the same multi-protein complex also show a differential turnover rate in one of the subcellular compartments. For example, Sm proteins, (components of the small nuclear ribonucleoprotein (snRNP) spliceosome subunits), showed faster turnover rates of ~18 hours in the cytoplasm, where snRNP proteins are assembled on snRNAs, compared with an average of ~35 hours in the nucleus, where the snRNPs function to splice pre-mRNAs (see Figure 43B and C). Interestingly, the Sm C subunit, which is not part of the same Sm complex as the other Sm subunits, did not show this difference in protein turnover between compartments. Other complexes with differences in subunit turnover rates between compartments include the 26S proteasome, nuclear pore, T-complex and RNA polymerase II. A common feature is that protein subunits have a faster turnover rate in the compartment where the complex assembles and are more stable in the compartment where the fully assembled complex functions.

6.3.5 Protein Characteristics Related to Turnover Rate

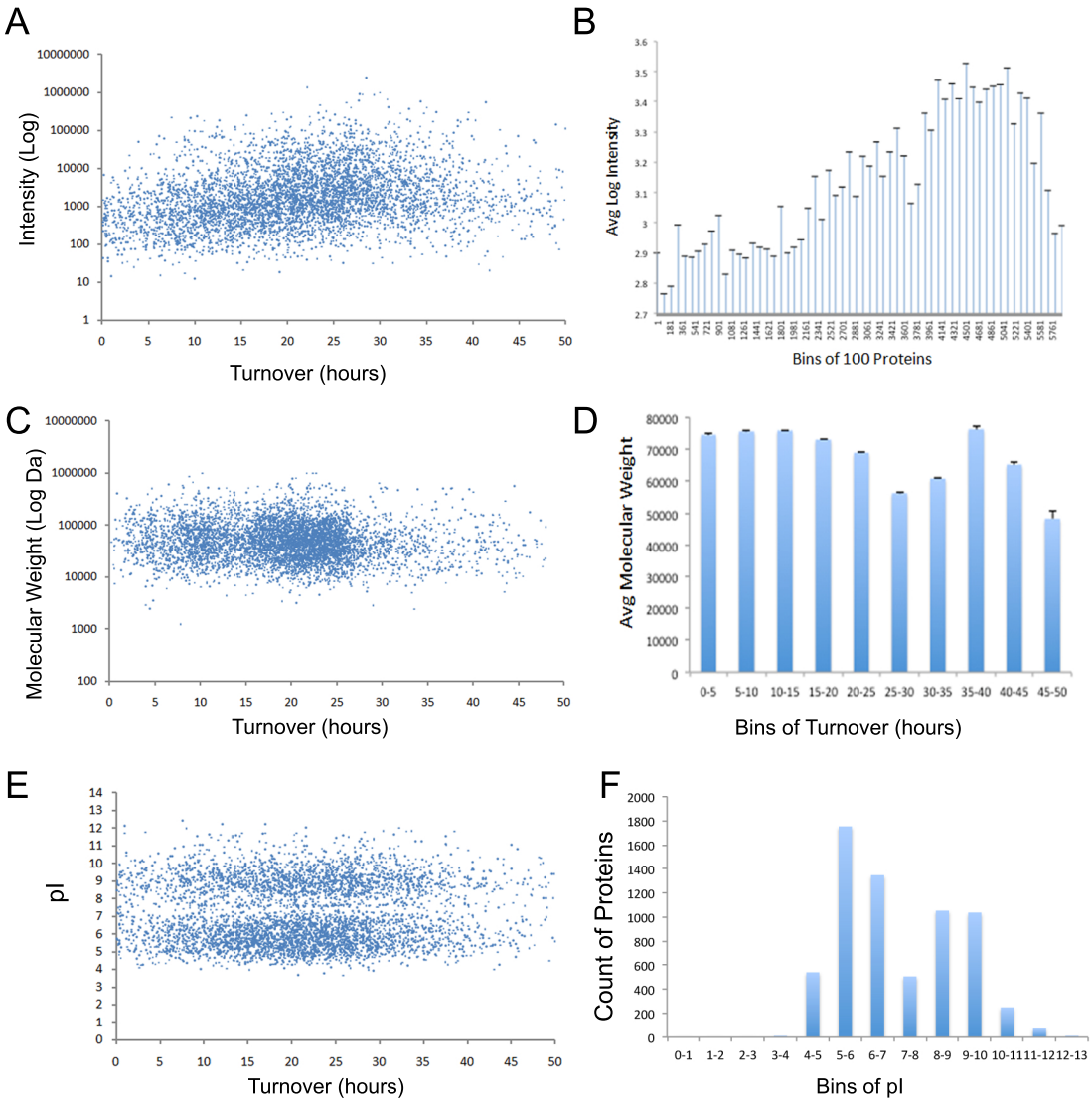


Figure 44: Protein characteristics related to turnover rate.

A) Protein abundance was estimated from the averaged sum of ion intensities measured for every peptide in a protein and plotted on the y-axis versus the turnover on the x-axis. B) A distribution plot with the average log base 10 intensity on the y-axis and bins of 100 proteins on the x-axis, where proteins are sorted from the fastest turnover to the slowest turnover for the whole cell. C) The log base 10 of molecular weight (in Daltons) was plotted versus the protein turnover in the whole cell. D) A distribution plot of the average molecular weight in Daltons on the y-axis and turnover (shown in 5 hour bins) on the x-axis. E) A comparison of the protein turnover on the x-axis with isoelectric point on the y-axis. F) A distribution plot of the number of proteins in each bin of isoelectric points.

A range of protein properties and characteristics, including abundance, size, pI values, sequence motifs and amino acid composition were analysed for correlations with turnover rates. A positive correlation was detected between protein abundance, as measured from peptide intensity, and the rate of protein turnover (see Figure 44A and

B). This correlation was also recently observed in a study of protein turnover in mouse cells (Schwanhausser et al., 2011). While there is variation, higher abundance proteins generally had slower than average turnover rates (see Figure 44B). The corollary is that the time to turn over half of the total protein content of a HeLa cell is ~ 15-20% longer than the mean turnover value of all proteins measured.

In contrast with the positive correlation with abundance, there is minimal correlation between turnover rate and protein size (see Figure 44C and D). Comparison of predicted molecular weights deduced from amino acid sequences with measured protein turnover rates showed a Pearson correlation coefficient of -0.09 (see Figure 44B). It has been proposed that acidic proteins are degraded more rapidly than basic proteins (Dice and Goldberg, 1975). A comparison of the rate of protein turnover with isoelectric point however showed no correlation (see Figure 44E, Pearson correlation 0.009). It is concluded that the bulk charge property of proteins is not a significant determinant of their stability. However, the analysis showed that nucleolar proteins have an inverse correlation between pI and protein turnover, with a Pearson correlation of -0.23. Thus, basic nucleolar proteins have a faster than average turnover (see Figure 42E, purple), likely due to the large number of basic ribosomal proteins in the nucleolus, which have a very fast turnover.

The presence of protein segments rich in proline, glutamic acid, serine and threonine, (PEST sequences) are reported to affect degradation levels (Rogers et al., 1986). Therefore, the proteins whose turnover was measured for the presence and frequency of PEST motifs were analysed using the PEST-find tool (http://emboss.sourceforge.net/). Parameters used include a minimal distance of 10 positively charged amino acids and a threshold of 5, to differentiate between weak and strong potential PEST motifs. All PEST motifs assessed as ‘poor’ and ‘invalid’ were removed from the analysis. While no simple correlation was observed between the presence of PEST sequences per se, and short half-lives, the average number of PEST regions found in proteins with a shorter than average turnover was ~1, as compared with a PEST frequency of ~0.5 for proteins with a longer than average turnover. It is concluded that there is a positive relationship between the presence of PEST motifs and protein turnover rates in HeLa cells but PEST motifs alone are not solely

responsible for rapid protein turnover. The data suggest that PEST motifs are only one of multiple factors that can affect protein half-lives.

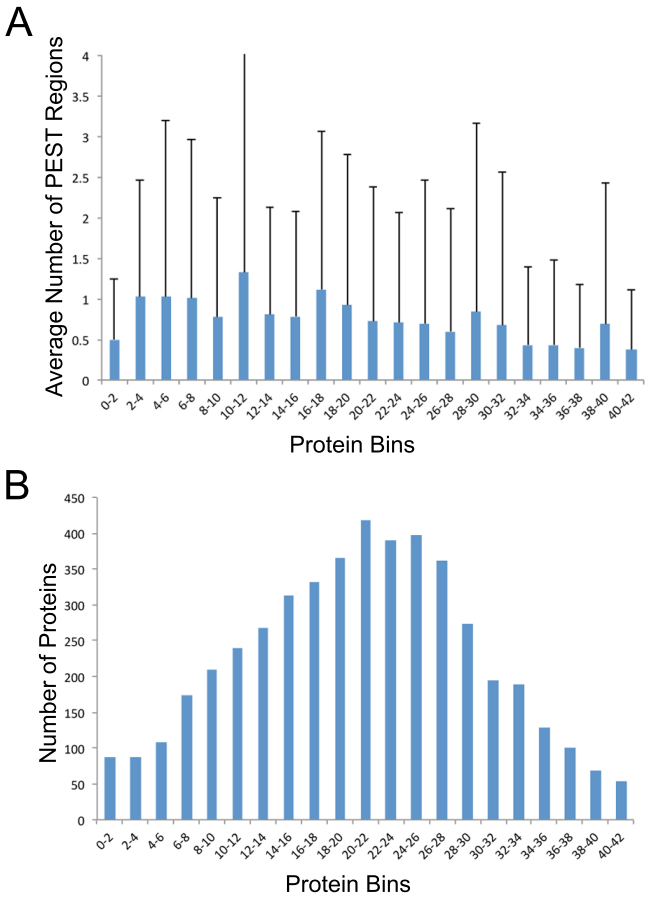


Figure 45: PEST sequence analysis.

A and B show all proteins binned in 2 hour segments, based on their respective turnover values (X axis). Axis Y in A shows the average number of PEST regions for the proteins in each bin and B shows how many proteins are in each bin. As expected, B shows a larger number of proteins in the bins near the average turnover found for the human proteome in this study (i.e. ~20 hours). However, as A suggests, there is no simple correlation between the turnover rate of a protein and the number of PEST regions found in the protein sequence.

6.3.6 Protein Turnover and the N-terminal Amino Acid Rule

A previously characterised determinant of protein stability is the N-terminal amino acid of the mature protein, where the N-terminal amino acid is classified as either stabilizing, or destabilizing (Varshavsky, 1992). While for most proteins, methionine is the first amino acid encoded and translated, methionine aminopeptidase is thought to remove the N-terminal methionine when the second amino acid is either C, G, A, S, T,

C or P (Hu et al., 2006). Some mature proteins can also be generated by post-translational cleavage, resulting in different amino acids occurring at the N-terminus. The empirical measurements of protein turnover rates reported were used to test whether the identity of the N- or C-terminal amino acid could affect endogenous protein stability in HeLa cells. The turnover rates were averaged for all proteins measured with each amino acid at either the first ten N-terminal positions (i.e. +1 to +10), or the last ten positions from the C-terminus. No significant differences were observed between mean protein turnover rates according to the amino acid identity at either the first or last ten N-terminal, or C-terminal positions, respectively. This comparison was also made specifically for the 10% fastest and 10% slowest turnover proteins and also saw no correlation with amino acid identity at N- or C- termini. It is concluded that protein stability for full-length, endogenous proteins in HeLa cells is not determined primarily by either N-terminal or C-terminal rules based upon amino acid identity.

6.3.7 Amino Acid Frequency Distribution

The matrix of frequencies for each amino acid occurring at either the first ten N-terminal positions or last ten C-terminal positions in each protein identified was also determined and compared with the corresponding in silico prediction of amino acid frequencies for each ORF in the human genome. The resulting Pearson correlation coefficient of 0.99 shows that the sample of 6,402 proteins for which turnover rates were measured have a near identical distribution of N- and C-terminal amino acid frequencies to the total translated human proteome. It is concluded that the subset of human proteins sampled in this study is thus highly representative of the total human proteome.

6.3.8 Turnover Viewer

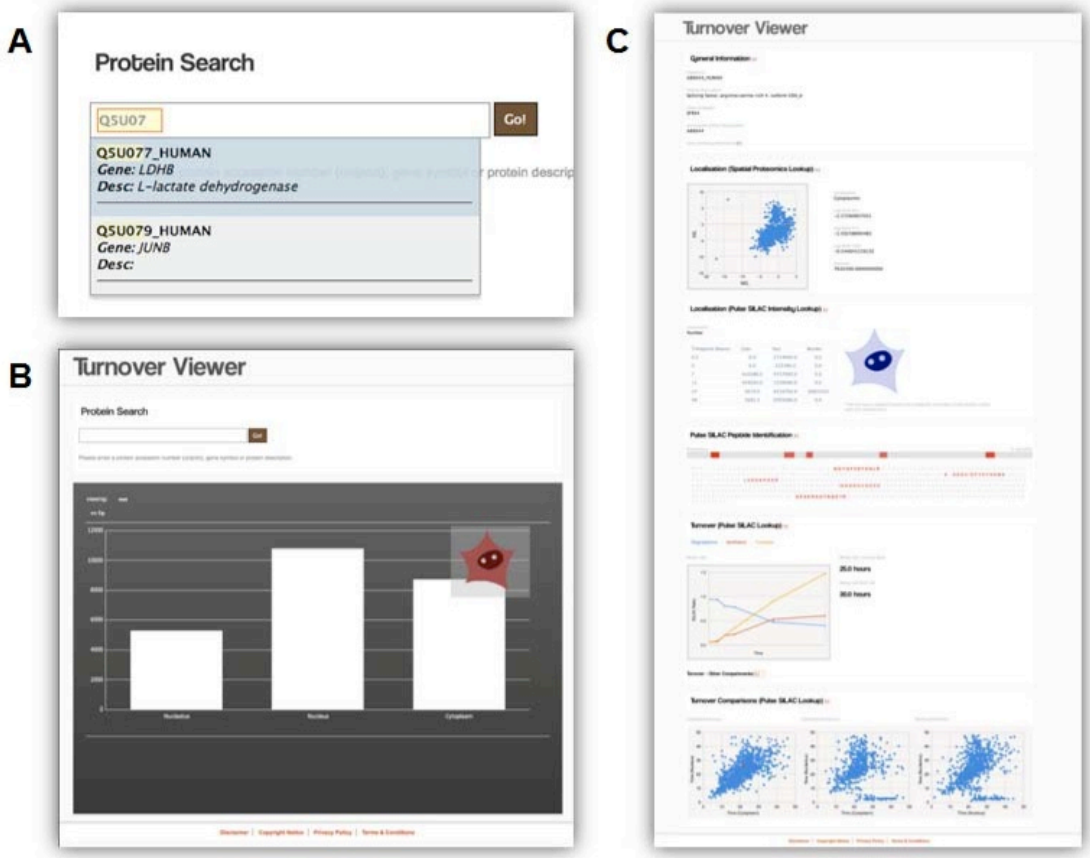


Figure 46: PepTracker spatial and turnover viewer.

A database viewer has been setup to provide access to the data generated during this study through a web-based interface (http://peptracker.com/turnover/). The application includes a search facility (A) that allows users to search for a protein(s) of interest using protein name, description, gene as well as IPI or Uniprot identifiers. A user can also select a protein using the interactive chart component that is provided on the home page (B). A more detailed page describing specific proteins (C) which documents previous subcellular localisation data from human HCT116 cells, calculated using spatial proteomics, as well as the localisation data resulting from averaging the intensity of the peptides identified in each subcellular compartment in this study. The viewer displays the peptides identified over the protein sequence, with different shading of red for each peptide reflecting differences in turnover rates. The different curve fits are displayed showing the degradation and synthesis rates for each protein, and showing the turnover rate of each protein in each subcellular compartment. The protein turnover for the selected protein is shown overlaid in red on the scatter plots showing all of the proteins in the different subcellular compartments in blue.

A database viewer has been implemented within the PepTracker software environment to provide convenient access to these data through a web-based interface (see Figure 46) (http://www.peptracker.com/turnover/). The application includes a search facility that allows users to search for a protein(s) of interest using protein name, description, gene as well as IPI or UniProt identifiers. The search result

page (see Figure 46C) documents also spatial proteomics localisation data from human HCT116 cells (Boisvert et al., 2010), as well as the localisation data resulting from averaging the intensity values of the peptides identified in each subcellular compartment in this study. The viewer displays all peptides identified for a selected protein sequence, with different shading for each peptide reflecting differences in turnover rates. The curve fits are displayed showing degradation and synthesis rates for each protein, and showing the turnover rate of each protein in each subcellular compartment. Protein turnover for a selected protein is shown overlaid in red on scatter plots showing all proteins in the respective subcellular compartments in blue, providing a rapid overview of turnover rates between subcellular compartments.

6.4 Technical Implementation

The PepTracker turnover and spatial viewer consists of a web-based, multi-tier architecture, where the data storage, server-side logic and user interface are separate components. The data storage is implemented as a fully relational Oracle database (Oracle Database 10g Enterprise Edition Release 10.2.0.5.0). This database holds turnover details, both at the protein and peptide level.

The server-side logic and client interface reside on an Apache web server (Version 2.2.3 - CentOS Linux Distribution). The server-side logic is built as an extension to the main PepTracker software. It is implemented using Python (Version 2.6.4), structured using the Django framework (Version 1.2.1, http://www.djangoproject.com/). The Django framework enforces code to follow the model view controller (MVC) design pattern, thus the functionality within the application is separated from the overall look and feel of the application, ensuring a more customisable solution. The server-side logic makes use of complex Structured Query Language (SQL) in order to communicate with the database and extract the relevant data required by the user interface. It does so using SQLAlchemy (http://www.sqlalchemy.org/), a Python SQL toolkit and object relational mapper. The Django framework provides the ability to setup html templates that form the user interface. These templates are customised on the fly using the data passed to them from the server-side logic. The templates are coded in HyperText Markup Languages (HTML), with Javascript additions to enhance functionality. Specifically, the viewer makes use of the JQuery library (http://jquery.com/) and Google Visualisation API

(http://code.google.com/apis/visualization/interactive_charts.html) to provide additional elements, such as interactive charts. The interface also includes an Adobe Flex component that provides an interactive chart and cell map for users to navigate through the data. To further enhance the user experience the user interface performs dynamic requests to the server using REST (Representational State Transfer). These requests prevent HTML pages from having to be completely redrawn and instead only the relevant sections of a page are updated.

6.5 Discussion

A combined pulse-labelling, spatial proteomics and data analysis strategy was developed to characterize the expression, localisation, synthesis, degradation and turnover rates of endogenously expressed, untagged human proteins in different subcellular compartments. Using SILAC and mass spectrometry, a total of 80,098 peptides from an estimated 8,041 HeLa proteins were quantified, and their spatial distribution between the cytoplasm, nucleus and nucleolus determined and visualised using PepTracker. Using information from ion intensities and rates of change in isotope ratios, protein abundance levels and protein synthesis, degradation and turnover rates were calculated for the whole cell and for the respective cytoplasmic, nuclear and nucleolar compartments. Based on the large number of proteins quantified and the close correlation between the N- and C-terminal amino acid frequencies of the experimentally identified proteins and the in silico predicted frequencies based on translation of human genome ORFs, it is argued that the dataset reported here is highly representative of the properties of the human proteome.

This study provides the first systematic, system wide quantitative analysis of proteome localisation and turnover that has evaluated the properties of endogenous proteins in different subcellular compartments. The approach described, together with the software tools for data visualisation and analysis, provides a basis for further systematic proteome-wide characterization of protein localisation and turnover that can be compared between different cell types, cell cycle stages, physiological conditions and genetic backgrounds.

Many large-scale, functional genomics studies characterise global gene expression levels, either in different cell types and/or under a range of growth conditions, by measuring differences in mRNA expression levels. This either involves microarray

technology, or, more recently, high-throughput RNA sequencing. However, quantitative mRNA expression data alone are not sufficient to reliably document gene expression at the proteome level. Previous large-scale analyses correlating mRNA and protein expression have found that similar mRNA expression levels can be accompanied by a wide range (up to 20-fold difference) in the corresponding abundance levels of the proteins encoded by these mRNAs (Gygi et al., 1999b). In agreement, only weak correlations are observed (Pearson correlation coefficients ~0.2) comparing the estimates of HeLa protein expression levels from this study with publicly available HeLa mRNA expression data (ArrayExpress, EBI). This overall poor correlation between cognate mRNA and protein expression levels likely reflects differences both in rates of mRNA translation and in protein turnover, but may also result, at least in part, from noisy microarray measurements and/or variability of HeLa cell batches. It is important to note that many proteins can differ substantially in their in vivo half-lives, regardless of how fast they are synthesised (Greenbaum et al., 2003). This underlines the importance of making direct measurements of endogenous cell proteins, including the high-throughput analysis of protein turnover, to fully evaluate gene expression responses and accurately determine factors and mechanisms regulating intracellular protein abundance.

Together with Lamond Laboratory collaborators, previously a heavy–light amino acid pulse SILAC protocol was used to measure protein turnover in HeLa cell nucleoli and this study also compared the MS data with parallel studies on turnover of GFP-tagged nucleolar proteins using fluorescence microscopy (Lam et al., 2007). A similar heavy – light pulse SILAC approach was recently used to study turnover of human A549 lung carcinoma cell proteins (Doherty et al., 2009) and also to study mouse NIH3T3 cells (Schwanhausser et al., 2011). In my thesis work, the pulse SILAC technique is extended by analysing a combination of heavy – medium pulsed cells and an equal amount of control, light cells at each time point. This offers advantages in terms of improved data analysis and statistical evaluation procedures.

Overall, the pulse SILAC approach is useful for determining protein turnover because it allows measurement of endogenous proteins expressed at physiological levels while avoiding the need to treat cells with translation inhibitors. Techniques based on translation inhibition complicate the interpretation of protein turnover values as the

effect of the inhibitors on cell physiology, which can in turn affect protein stability, must be taken into account (Belle et al., 2006). The pulse SILAC method, as previously demonstrated, can also be used to compare the turnover rates of tagged and endogenous forms of the same protein in stable cell lines, which is often important to validate independently the conclusions to be drawn from microscopy studies in mammalian cells using GFP-fusion proteins (Lam et al., 2007).

Several lines of evidence argue that the MS-based proteomics approach used here is robust and reproducible. First, a strong positive correlation (Pearson correlation coefficient ~0.73) was observed between the present data and the smaller dataset from previous pulse SILAC analysis of HeLa nucleolar protein turnover (Lam et al., 2007), despite the lower number of peptides identified for each protein in that case. This demonstrates that, at least when comparing the same cell line, biological replicates produce similar results. Second, it has been confirmed that values from technical replicates are reproducible by evaluating data obtained from separate MS analysis and quantitation of the same protein samples using different mass spectrometers. Thirdly, it is noted that there was a strong positive correlation between the localisation and turnover values obtained for most shared subunits of common multi-protein complexes, consistent with proteins in the same complex having similar biological properties. Most peptides in the same protein also produced similar values.

This approach differs in several aspects from most previous studies on global protein turnover (Belle et al., 2006, Yen et al., 2008, Doherty et al., 2009, Eden et al., 2011, Schwanhausser et al., 2011). First, it was decided to simultaneously determine not only net protein turnover, but also both protein degradation and synthesis rates. This provides additional information on the protein properties and allows calculation of turnover using two separate methods and statistical evaluation of the accuracy of the turnover values for each protein. Second, protein turnover was characterised not only for the global protein population in whole cells, but also for proteins in separate subcellular fractions. This spatial information recognises that separate pools of protein with distinct properties can exist in different subcellular locations. Inevitably, analysis of whole cell extracts measure ensemble, average values for the protein population and will not identify cases where the same protein can be present in more than one complex with different turnover rates, as revealed in this study. Third, measurements

are made on endogenous, untagged cell proteins and not based upon analysis of over-expressed, tagged fusion constructs. Either transient, or stable, overexpression of tagged fusion proteins may affect their turnover properties, both through changing the protein structure and by altering their abundance and stoichiometry relative to interaction partners.

A comparison of the protein turnover values reported here with the corresponding turnover or half-life values reported for the same proteins in previous large-scale studies showed major differences. Thus, there appeared to be a near random correlation with the values reported in two of these studies (Doherty et al., 2009, Eden et al., 2011), and only a partial positive correlation (Pearson correlation coefficient ~0.2) with the data of Yen et al. (Yen et al., 2008). However, it is noted that cross comparison of the datasets from each of these previous studies also shows mostly random correlations between them. However, we found a stronger correlation (Pearson correlation coefficient of 0.34) between our data and a recent study also using a pulse-SILAC method to analyse protein turnover in mouse NIH 3T3 cells (Schwanhausser et al., 2011).

The lack of correlation between the previous high-throughput studies and this present data is not simply explained by variation in the quality of the data in this study. Even focusing on the subset of proteins in this study with the highest quality measurements did not significantly improve the degree of correlation with the previously published analyses. Thus, considering only the ~25% of proteins for which this study quantitated at least 20 separate peptides with optimal chi^2 curve fitting did not increase the positive correlation with the other datasets. It is concluded that our study has generated a dataset for endogenous human proteins that is distinct from previous studies and, considering the overall stringent data evaluation employed, argue that the lack of agreement in protein turnover values between this data and previous large-scale studies is not primarily reflecting data quality issues in measurements described here.

This surprising situation where apparently each separate high-throughput analysis of protein turnover produces different results could have multiple explanations. Two of the previous studies specifically analysed the turnover of over-expressed, GFP-tagged fusion proteins (Yen et al., 2008, Eden et al., 2011). It is anticipated that the resulting

fusion protein turnover values may differ from the rates of turnover reported here for the corresponding endogenous proteins expressed at physiological levels. It is also important to note, however, that there is no expectation that different cell lines growing under different culture conditions should show identical protein turnover values, as observed from the differences in turnover measured between A549 cells (Doherty et al., 2009), NIH 3T3 cells (Schwanhausser et al., 2011) and HeLa cells (this study). It will be interesting, therefore, to carefully evaluate differences in protein turnover rates between cell lines and growth conditions using the same stringent methodologies for all measurements. Indeed, while most homologous proteins showed a similar trend when comparing turnover values from our study to those found in NIH 3T3 cells (Schwanhausser et al., 2011), it is interesting to note that a few proteins showed a dramatic difference in protein turnover between the two cell lines, indicating that specific protein degradation might be either cell type specific, or species-specific, or both.

An interesting general feature from this study is the observation that protein subunits from multi-protein complexes show faster turnover as free proteins, prior to complex assembly. This is exemplified by the ribosomal proteins, which have a 50% turnover of ~6 hours in the nucleus, where the protein pool includes free, unassembled ribosomal proteins (Lam et al., 2007). In contrast, ribosomal proteins are very stable in the cytoplasm, with turnover of over 30 hours, where they accumulate only after assembly into a ribosome subunit. An important corollary is that a large increase in the expression of a specific protein, as often occurs upon either transient or stable over-expression of tagged proteins, may change drastically its turnover in comparison with the endogenous counterpart. The measured turnover of interaction partners of the over-expressed factor and of other proteins may also be altered. It is proposed that this could account for much of the difference between the turnover values reported here for endogenous proteins expressed at physiological levels and the faster turnover rates measured using fluorescent protein-tagged proteins. For example, Yen et al., reported that the turnover rates of over 8,000 GFP-tagged human proteins showed a bimodal pattern, where the average turnover values were measured as 30 min and 2 hours, respectively (Yen et al., 2008). This study also found that protein turnover followed a bimodal distribution, but with slower turnover values of ~20 hours, close to the HeLa cell division rate, and a minor peak with a turnover value of ~10 hours.

The data indicate that the bulk of HeLa cell proteins may be turned over passively during normal cell growth and are consistent with the mean turnover rate reflecting approximate doubling of the amount of proteins as the cell divides and hence doubles its protein content. However, a subset of proteins show faster turnover, suggesting they may be directly targeted for degradation. The similar distribution of protein turnover rates seen for the cytoplasm and the nucleoplasm likely reflects the high level of protein shuttling between these compartments. This contrasts with the nucleolus, where a distinct group of proteins show fast turnover (<6 hours), mostly corresponding to ribosomal proteins. Interestingly, recent studies point to a role for the accumulation of specific free ribosomal proteins in the nucleus in signalling mechanisms involved in stress responses and growth control (Sundqvist et al., 2009), suggesting that the control of ribosomal protein stability in the nucleus is involved in biological regulation.

Proteins with the slowest turnover have a wide range of functions, but are commonly present either in large, abundant and stable protein complexes, such as ribosome and spliceosome subunits, RNA polymerases, the nuclear pore, the exosome and the proteasome, or else are found inside mitochondria. Interestingly, with almost all of these slow turnover proteins, it is noted that the turnover rate of each subunit was significantly slower in one subcellular compartment, correlating with the location where they exert their function. These observations suggest a general assembly strategy whereby cells produce an excess of subunits in order to favour complex formation, but carry out this assembly in a compartment separate to the eventual main site of function. This avoids the need to tightly co-regulate transcription, processing, transport and translation of the mRNAs encoding different protein subunits in eukaryotes where genes are not organised in operons and not co-transcribed and translated. Any excess protein subunits produced will simply be degraded in the assembly compartment. This model explains the differential stability of ribosomal proteins between the nucleus, where they are assembled with RNA, and the cytoplasm where they function to translate mRNA and conversely, the higher stability of snRNP proteins in the nucleus, where they function in pre-mRNA splicing, as opposed to in the cytoplasm, where they assemble on snRNAs.

The empirical measurements of turnover values allows an objective evaluation of protein features that influence stability for endogenous cell proteins and a parallel

analysis of how this may vary between subcellular compartments. Thus, it is observed a general correlation that highly abundant proteins are more stable than average, consistent with the enduring roles of abundant structural proteins and major complexes involved in gene expression. In contrast, the net charge and molecular weight of proteins shows little or no correlation with turnover rate. Somewhat surprisingly, given the previous literature describing an 'N-end rule', whereby protein degradation mechanisms favour more rapid degradation of protein and peptide fragments with specific amino acids at the amino terminus, no evidence was found for either an N-end or C-end rule for endogenous HeLa proteins. Thus, the mean turnover rates for HeLa proteins is remarkably similar regardless of which amino acid is present at either the first ten, or last ten, positions in the polypeptide chain. This contrasts with budding yeast, where a dramatic difference in stability is reported for proteins with either Arg, Lys, Asp, Phe or Leu at their N-terminus, which show rapid degradation, as opposed to proteins having Met, Gly, Ala, Ser, Thr or Val at their amino terminus, which have half-lives significantly longer than the division rate of yeast cells (Varshavsky, 1992). Nonetheless, currently it can not be excluded that a more complex sequence motif pattern at the terminus of human proteins might correlate with degradation efficiency and it is also possible that the degradation rate of cleaved protein fragments, rather than full length proteins, may be affected in HeLa cells by the identity of specific amino-terminal amino acids or motifs.

A positive correlation is observed between the prevalence of PEST sequence motifs and rapid protein turnover, consistent with prior evidence that PEST motifs destabilized proteins in vivo (Rechsteiner and Rogers, 1996). However, the considerable variability in the turnover rates of specific proteins containing PEST motifs pointed to a complex relationship between sequence motifs and protein stability. It is likely that multiple structural and sequence elements, as well as abundance, localisation and the presence of interaction partners, can all affect the net stability of individual proteins. It is thus difficult to accurately predict protein stability based on primary sequence information and this demonstrates the importance of determining protein turnover empirically under different cell growth conditions. It is anticipated that at least some patterns of post-translational modification will also be found to correlate with protein properties such as turnover rate, subcellular localisation and abundance. Future work will address this issue and mining of

relational databases of protein properties and PTM patterns may predict functional relationships that can be evaluated further experimentally.

It is envisaged in future that this general approach for characterising protein turnover rates and associated protein properties such as subcellular localisation, abundance, interaction partners and PTM patterns can be extended in several ways. It is possible to expand the subcellular fractionation strategy for example and thereby obtain higher resolution spatial information regarding the subcellular distribution of the proteome and how this correlates with protein structure, isoforms and PTM patterns. This present study has not distinguished effects on the proteome of cells growing at different stages of the cell cycle. However, specific examples are already known where either protein stability or subcellular localisation can alter as cells progress through interphase and mitosis. Work is in progress therefore to carry out systematic, proteome wide analyses of how protein properties, including turnover rates and subcellular localisation patterns, vary as a function of cell cycle progression, providing a detailed quantitative annotation of the human proteome in both time and space. None of the protein properties discussed above represent 'absolute' values, and it is to be expected that rates of protein turnover, localisation patterns, interaction partners and PTMs will vary considerably between different cell lines, under different growth conditions and in response to drugs or other external stimuli. Specific mutations, which may be associated with either oncogenic transformation or genetic disease, can also alter these protein properties. The development and integration of many large-scale, quantitative proteomic datasets of the sort described here thus offers a promising future direction for expanding the functional annotation of the human genome, and the genomes of other model organisms, and for the discovery of new biological regulatory mechanisms.

6.6 Distribution of Effort

Lamond Laboratory biologists, primarily Francois-Michel Boisvert, carried out the experimental bench work during this study. The analysis described in this chapter was a joint effort between Yasmeen Ahmad and Francois Michel Boisvert. Software, created by Marek Gierlinkski, was used to aid with the data analysis, i.e. to build smoothed synthesis, degradation, turnover and intensity profile curves for each

protein identified by MS. Yasmeen Ahmad carried out all of the technical implementation with regards to the Turnover & Spatial Viewer.

Chapter 7: Protein Isoform, Localisation and Turnover Analysis

7.1 Summary

In higher eukaryotes, many genes encode two or more protein isoforms and the properties and biological roles of these separate isoforms are often poorly characterised. This chapter describes a number of systematic approaches developed for the detection of protein isoforms with differential biological properties (Ahmad et al., 2011).

Previously, information from ion intensities and rates of change in SILAC isotope ratios allowed calculation of protein abundance levels, turnover rates and subcellular distribution between cytoplasmic, nuclear and nucleolar compartments for the HeLa cell proteome (see Chapter 6: Spatial Localisation & Turnover Analyses). Protein isoforms were detected using three data analysis strategies: candidate approach, rule of thirds approach and three in a row approach. These strategies evaluate differences between SILAC isotope ratios for specific groups of peptides within the total set of peptides assigned to a protein. For known isoforms, the candidate approach compares SILAC isotope ratios for peptides predicted to be isoform-specific, with ratio values for peptides shared by all the isoforms. The rule of thirds approach compares the mean isotope ratio values for all peptides in each of three equal segments along the linear length of the protein, assessing differences between segment values. The three in a row approach compares the mean isotope ratio values for each sequential group of three adjacent peptides, assessing differences with the mean value for all peptides assigned to the protein.

Protein isoforms were also detected and their properties evaluated by fractionating cell extracts on 1-D SDS PAGE prior to trypsin digestion and MS analysis and independently evaluating isotope ratio values for the same peptides isolated from different gel slices. This strategy allowed detection of isoforms that migrate across multiple gel slices. Furthermore, the effect of protein phosphorylation on turnover rates was analysed individually for cytoplasmic, nuclear and nucleolar compartments by comparing the mean turnover values calculated for all peptides assigned to a protein, either including, or excluding, values for cognate phosphopeptides. This

showed that phosphorylation affected turnover of nucleolar proteins to a greater extent than for proteins localised to the nucleus or cytoplasm. Collectively, these experimental and analytical approaches provide a framework for expanding the functional annotation of the genome.

Chapter 7 describes the analysis protein properties to identify protein isoforms, focusing first on the occurrence of protein isoforms (section 7.2), following with a description of the bioinformatics analysis tools used (section 7.3), description of the novel approaches employed (sections 7.4) and finally a discussion on the importance of protein properties in data analysis (section 7.5).

7.2 Background

Biological regulatory mechanisms and cellular responses are predominantly mediated by proteins and multi-protein complexes. The structures and properties of these proteins are crucial for their function and can vary greatly. For example, protein expression levels in mammalian cells vary over a large dynamic range of 10^6 or more (Corthals et al., 2000), while subcellular localisation patterns, post-translational modifications, rates of synthesis and degradation and interactions with partner proteins are also variable properties (Hinkson and Elias, 2011). Furthermore, all of these properties not only vary between proteins, they are also dynamic and can vary for the same protein at different times, depending on parameters such as cell cycle progression, growth rate and signalling events. Proteomes are thus inherently complex and their properties in constant flux. This presents a major challenge for proteomic studies, which ideally should not only identify which proteins that are expressed in a cell or organelle, but also characterise their properties and quantify how these change in response to different perturbations and cell cycle stages etc. (Yates et al., 2005).

In higher eukaryotes, proteomic studies are complicated further by the fact that many genes encode two or more separate protein isoforms (Jungblut et al., 2008). Alternative splicing of pre-mRNA transcripts is commonplace and this can generate multiple mRNAs from the same gene and hence multiple different proteins (Matlin et al., 2005). Such isoforms can vary in length, share common exons, include variable exons and even have very different amino acid sequences because splicing events can alter the translational reading frame of the differentially spliced mRNAs. Isoforms can also arise from differential post-translational processing and modification of a

polypeptide encoded by a single mRNA. In other cases gene duplication results in expression of closely related protein paralogs that share extensive sequence identity and are hard to distinguish. Even minor structural differences between isoforms can alter their biological properties and result in distinct pools of related proteins whose subcellular location, function and interactions vary.

The expression levels, structures, properties and biological roles of separate protein isoforms are still poorly characterised and commonly ignored in many large-scale proteomic analyses (Jungblut et al., 2008). However, since protein isoforms usually share common sequences, even if one or more exons are isoform-specific, it is important to the biological interpretation of proteomics experiments to decide whether all peptides mapped to a specific gene are encoded in a single polypeptide. If instead the peptides derive from two or more isoforms, these may have distinct biochemical properties, such as subcellular localisation and/or turnover rate. In this case the averaged value from all of the peptides may give a misleading picture regarding the property of the protein under study. For example, when studying subcellular localisation, the averaged value for all peptides from a gene may indicate that the protein is present in both the cytoplasm and the nucleus, when in fact one isoform is predominantly cytoplasmic and the other predominantly nuclear (see Figure 47). This is likely to be of general importance for annotating the genome because a recent comparative study of subcellular protein localization in three human cell lines detected ~40% of the 4,000 genes analysed localising to multiple subcellular compartments (Fagerberg et al., 2011).

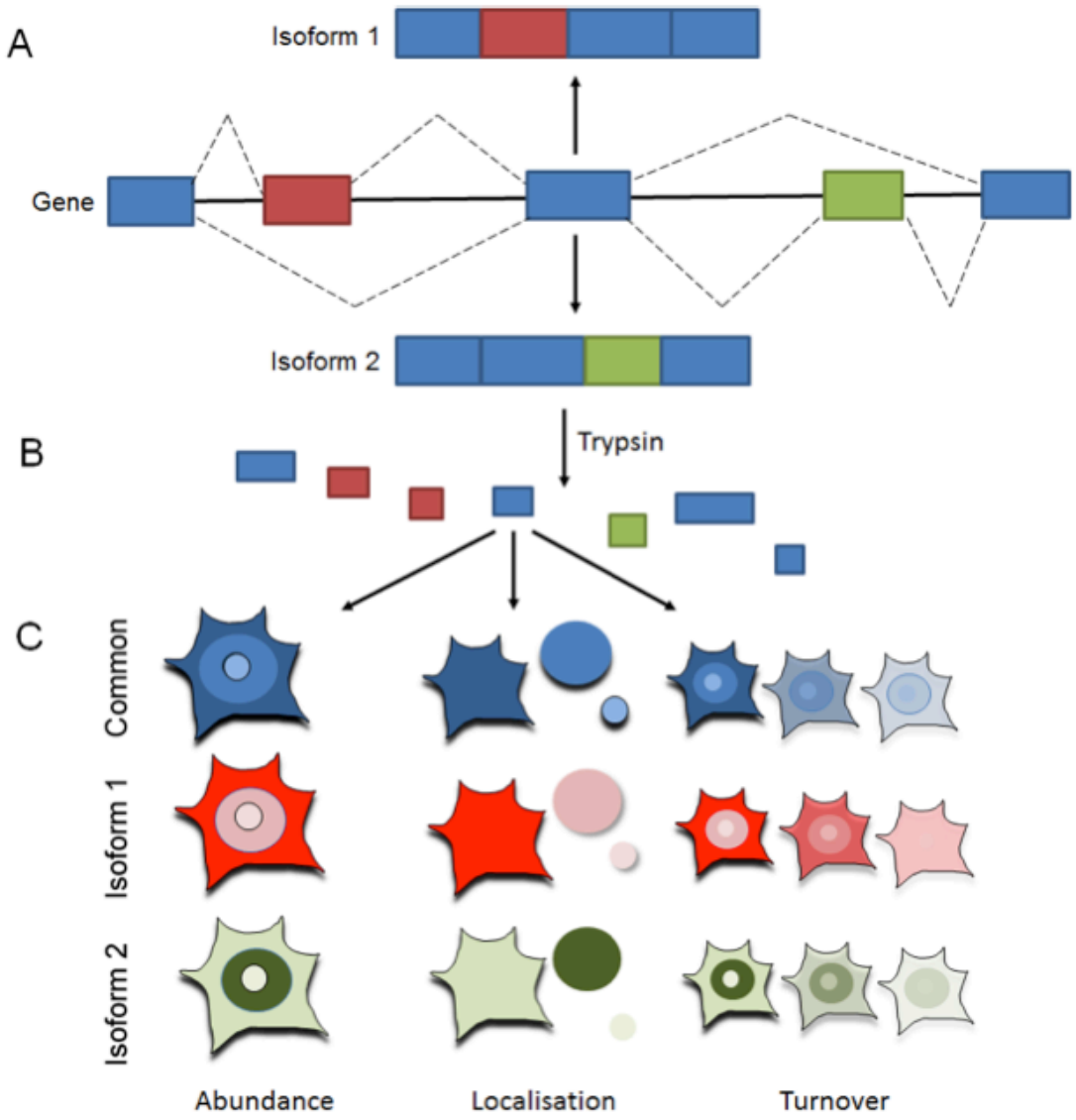


Figure 47: Alternative splicing leading to protein isoforms.

A) A single gene can encode multiple proteins due to alternative splicing. After transcription of a gene, exons of the resultant RNA can be reconnected in multiple ways during RNA splicing, resulting in the translation of protein isoforms. Shown are two isoforms produced from the same gene, with red and green areas signifying differences between the isoforms due to alternative splicing. B) Protein isoforms represent several different forms of a protein and have a largely shared sequence, however small differences occur. Protein isoforms can be recognised and differentiated via these differences, by locating isoform specific peptides, i.e. amino acid sequences that encode the difference. During mass spectrometry analysis, proteins are digested using an enzyme, often trypsin, which fragments a protein into peptides. A mass spectrometer can then be used to analyse samples and hence identify protein peptides. The list of identified peptides contains common peptides shared by all isoforms, but also specific peptides (shown in red and green) that are unique to each isoform. C) Using these identified peptides it is possible to extract abundance, localisation and turnover information for a protein. Using commonly shared peptides provides average abundance, localisation and turnover values for a protein, however, using isoform specific peptides it is possible to calculate values per isoform.

Mass spectrometry-based proteomics has become the technology of choice for the direct identification and characterisation of proteins (Walther and Mann, 2010a). In combination with quantitative approaches, such as SILAC (Stable Isotope Labelling with Amino acids in Cell culture), mass spectrometry can not only identify proteins and post-translational modifications, but also measure how relative protein levels change in cells under different conditions (Ong et al., 2003, Mann, 2006). This provides a flexible assay format for proteomic studies that evaluate differences between two or more cell states, each defined by metabolic labelling of proteins with amino acids that have different combinations of isotopes incorporated into selected amino acids. Subsequent isolation of proteins and enzyme cleavage results in mixtures of isotopically labelled peptides where the relative levels of each isotopic form can be resolved and quantified by mass spectrometry. The peptide isotope ratios are then mapped back to the genome sequences encoding the cognate proteins and used to infer whether either the levels, or properties, of these proteins have been changed. The SILAC strategy has been used for quantitative studies of cell and organelle proteomes and for comparative studies of protein modifications, and interactions (Walther and Mann, 2010b) and to identify proteins isolated from mitotic chromosomes (Ohta et al., 2010). It has also been used in combination with cell fractionation to generate 'isotope-encoded' subcellular compartments allowing subcellular protein localisation to be evaluated on a system-wide level (Boisvert et al., 2010).

Chapter 6 reports a global analysis of protein abundance, subcellular localisation and turnover in HeLa cells using SILAC and mass spectrometry that characterised over 80,000 peptides mapped to ~8,000 human genes (Boisvert et al., 2011). In this chapter this HeLa dataset was analysed using systematic approaches for the detection of protein isoforms with differential biological properties. Methods were evaluated that can identify human protein isoforms whose turnover and/or subcellular localisation properties vary and analyse phosphorylated peptides that are correlated with altered rates of protein turnover in the separate cytoplasmic, nuclear and nucleolar compartments.

7.3 Quantification and Bioinformatics Analysis

The methods used for preparation of SILAC labelled HeLa proteins from nuclear, nucleolar and cytoplasmic fractions, protein chromatography by SDS PAGE, trypsin digestion and mass spectrometry was described previously (Boisvert et al., 2011). Peptide identification, quantitation and phosphopeptide analysis was performed using MaxQuant version 1.1.1.14 (Cox and Mann, 2008, Cox et al., 2009). The derived peak list was searched using Andromeda as the database search engine for peptide identifications against the International Protein Index (IPI) human protein database (version 3.68) containing 89,422 proteins, to which 175 commonly observed contaminants and all the reversed sequences had been added. The initial mass tolerance was set to 7 p.p.m. and MS/MS mass tolerance was 0.5 Da. Enzyme was set to trypsin/p with 2 missed cleavages. Carbamidomethylation of cysteine was searched as a fixed modification, whereas N-acetyl protein, oxidation of methionine and phosphorylation of serine, threonine and tyrosine were searched as variable modifications. Identification was set to a false discovery rate of 1%. To achieve reliable identifications, all proteins were accepted based on the criteria that the number of forward hits in the database was at least 100-fold higher than the number of reverse database hits, thus resulting in a false discovery rate (FDR) of less than 1%. A minimum of 2 peptides were quantified for each protein. Data analysis was performed using the PepTracker software environment. Clustering analysis was performed using the software Cluster with complete linkage clustering and visualised using Treeview (http://rana.lbl.gov/EisenSoftware.htm) (Eisen et al., 1998).

7.4 Results

7.4.1 Protein Isoform Analysis: Candidate Approach

The data described in Chapter 6, i.e. HeLa cell SILAC data describing global protein abundance, localisation and turnover (Boisvert et al., 2011), has been analysed using three approaches to detect protein isoforms that have differential properties (see Figure 47). First, a candidate approach was used. For genes encoding known isoforms, average intensity values were compared for peptides shared between all isoforms with candidate, isoform-specific peptides (see Figure 48). This is illustrated for the NudCD1 protein, which has three reported isoforms (Yan et al., 2004). Using average values for all peptides detected that are common to the three isoforms (blue), there is similar

average peptide intensity in the cytoplasm and nucleus, with little signal in the nucleolus. However, while analysis of a peptide predicted to be specific to isoform 3 showed intensity in both cytoplasmic and nuclear compartments (green, ~3:2 cytoplasmic:nuclear), a peptide predicted to be specific for isoform 2 (red), instead showed exclusively cytoplasmic signal (see Figure 2A). In the case of isoform 1, it was not possible to identify an isoform-specific peptide that could be reliably detected. However, as there is strong overall peptide signal in the nucleus that cannot be accounted for by the intensities of either the isoform 2, or isoform 3-specific peptides, it can be inferred that isoform 1 is likely to be enriched in the nucleus.

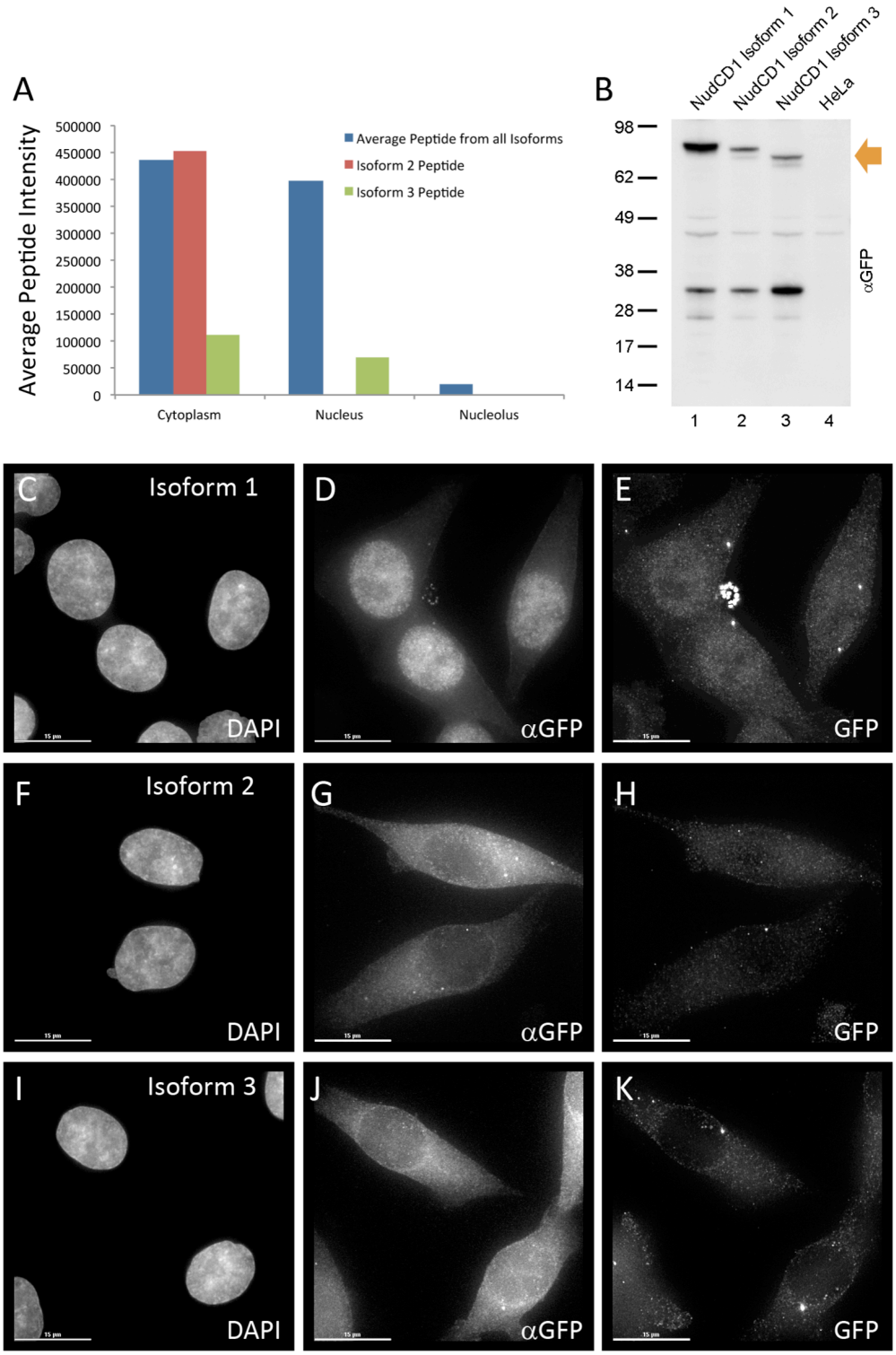


Figure 48: NUDCD1 Protein isoform identification and localisation.

This figure describes the identification of NudCD1, NudC domain-containing protein 1. A) Chart showing intensity (Y axis) in different cellular compartments (X axis) for three peptides. The blue series provides an average intensity of common peptides shared between all three known isoforms of NudCD1, indicating that the NudCD1 protein has approximately similar average peptide intensity in the cytoplasm and nucleus, with little signal in the nucleolus. The remaining series show the intensity of peptides not commonly shared by all three isoforms, i.e. MLYLQGWSMPAVAEVK (isoform 2 peptide

shown in red) and YNQDTALGKPR (isoform 3 peptide shown in green). The isoform 3 specific peptide shows intensity in both cytoplasmic and nuclear compartments (~3:2 cytoplasmic:nuclear), and the isoform 2 peptide shows exclusively cytoplasmic signal. B) HeLa cells expressing GFP fused at the N-terminus to isoform-specifc cDNAs were used to establish stable HeLa cell lines where expression of the fusion protein is under the control of a tetracycline-regulated promoter. All three stable HeLa cell lines produced proteins of the expected sizes when induced by addition of tetracycline and analysed by protein blotting, detected using an anti-GFP antibody. As can be seen by the upper bands on the western blot (orange arrow), three protein isoforms are recognised, migrating at the predicted molecular weights of NudCD1 isoforms, i.e. isoform 1 at 66.76kDa (lane 1), isoform 2 at 63.50kDa (lane 2) and isoform 3 at 56.61kDa (lane 3). Fluorescence microscopy analysis of HeLa cells expressing the respective GFP-fusion proteins was performed to determine localisation patterns, using both an antibody to GFP (panels D, G & J), and direct GFP fluorescence (panels E, H & K). NudCD1 Isoform 1 shows nuclear accumulation in panels D & E, NudCD1 isoform 2 shows predominantly cytoplasmic accumulation in panels G & H and both cytoplasmic and nuclear accumulation is shown of NudCD1 isoform 3 in panels J & K.

As no suitable isoform-specific antibodies for NudCD1 were available, the localisation patterns of the three NudCD1 isoforms were next compared by immunofluorescence microscopy analysis of HeLa cells expressing GFP fused at the N-terminus to isoform-specific cDNAs (see Figure 2, B-K). All three GFP-NudCD1 isoform fusions were used to establish stable HeLa cell lines where expression of the fusion protein is under the control of a tetracycline-regulated promoter. All three stable HeLa cell lines produced proteins of the expected sizes when induced by addition of tetracycline and analysed by protein blotting, detected using an anti-GFP antibody (see Figure 48B). Fluorescence microscopy analysis of HeLa cells expressing the respective GFP-fusion proteins was performed, using both an antibody to GFP (see Figure 48, panels D, G & J), and direct GFP fluorescence (see Figure 48, panels E, H & K), to determine their localisation patterns. In agreement with the spatial proteomics data, this showed predominantly nuclear accumulation of NudCD1 isoform 1 (panels D & E), predominantly cytoplasmic accumulation of NudCD1 isoform 2 (panels G & H) and both cytoplasmic and nuclear accumulation of NudCD1 isoform 3 (panels J & K). None of the three GFP-NudCD1 isoform fusions accumulated in nucleoli.

These data analysing NudCD1 isoforms illustrate the validity of the candidate approach but also highlight its limitations. It relies upon prior annotation to predict the existence of isoforms and the ability to detect unique peptides that are isoform-specific. As seen for isoform 1, it is not always possible to detect isoform-specific peptides. Even when isoform-specific peptides can be detected, as with isoforms 2 and 3, they are usually

few in number (often only one) and this reduces the accuracy of the overall quantitation. Nonetheless, the NudCD1 data show clearly that analysis of a key protein property, such as subcellular localisation, can be misleading when values for all peptides are averaged without taking into account the existence of distinct pools of protein with differential localisation phenotypes.

7.4.2 Protein Isoform Analysis: Rule of Thirds Approach

Next, two methods were used to systematically evaluate whether the mean value of all peptides quantitated for a given protein included clusters of adjacent peptides with significantly different mean values. First, a 'rule of thirds' approach was used to search the data for examples where the mean values of peptides from the amino terminal (S1, blue), central (S2, red) or carboxy terminal (S3, green) segments of the protein differed by at least one standard deviation from an adjacent segment. This was evaluated for over 6,000 HeLa proteins, where at least two peptides had been quantitated within each segment of the protein sequence. The mean turnover rate for each segmented third of every protein was plotted on the y axis against total proteins, ranked on the X axis by the mean turnover value derived from all peptides assigned to that protein (see Figure 49A). Examples where the turnover value of any one third segment of a given protein differed by more than 70% from the overall turnover value for the same protein, i.e. the mean of all the peptides assigned to that protein, are highlighted and colour coded in blue, red and green for segments S1-S3, respectively (see Figure 49A).

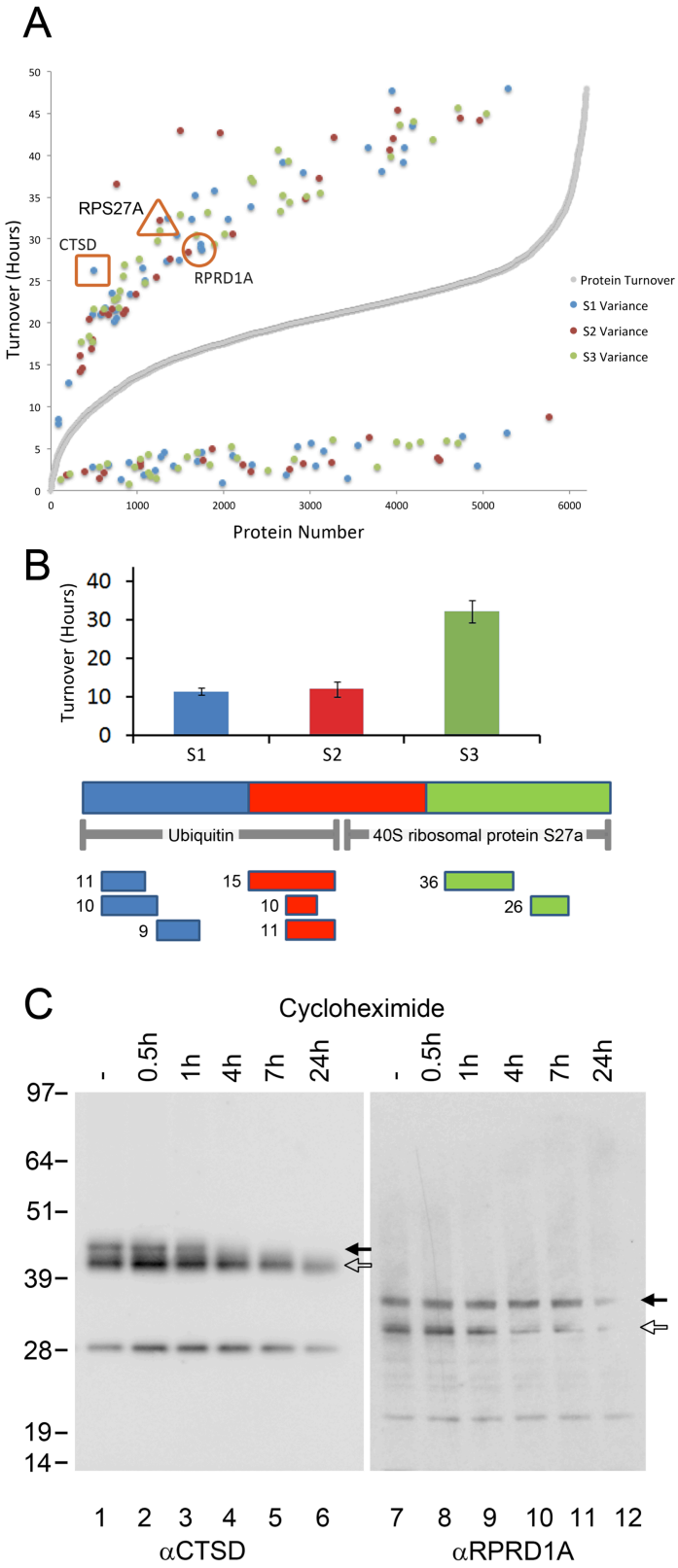


Figure 49: Protein isoform identification from protein sequence segmentation.

A) Graph showing all proteins (X axis) ordered by average turnover of all identified peptides against turnover in hours (Y axis). The grey series shows the average turnover for each protein, calculated from all peptides identified for the protein. To identify

potential isoforms, each protein was divided into three equal segments based on sequence length. A turnover value was calculated for each segment of the protein, using only peptides found in that segment. The blue, red and green series highlight segments of proteins that have a variance greater or less than 70% compared to the average protein turnover. RPS27A, Ubiquitin-40S ribosomal protein S27a, is a protein that shows a variance in segment 3 of the protein sequence (highlighted using orange triangle). B) The chart shows that the three equal segments of RPS27A (X axis), show an average turnover (Y axis) of 10.73 hours (segment 1), 14.46 hours (segment 2) and 31.06 hours (segment 3). As can be seen from the linear representation of the protein, the peptides for each segment (examples shown underneath) have a significantly different turnover in the carboxy terminal segment 3 (CCLTYCFNKPEDK: 26.49 hours, ECPSDECGAGVFMASHFDR: 35.63 hours) compared with segment 1 (EGIPPDQQR: 10.92 hours, IQDKEGIPPDQQR: 10.93 hours, LIFAGK: 12.02 hours, MQIFVK: 9.19 hours, TITLEVEPSDTIENVK: 10.62 hours, TITLEVEPSDTIENVKAK: 10.00 hours, QLEDGR: 11.41 hours) and segment 2 (ESTLHLVLR: 12.80 hours, QLEDGRTLSDYNIQK: 11.72 hours, QLEDGRTLSDYNIQKESTLHLVLR: 15.20 hours, TLSDYNIQK: 10.05 hours, TLSDYNIQKESTLHLVLR: 10.95 hours, SYTTPK: 26.07 hours). In fact, the full length RPS27A protein is expressed as a precursor that is subsequently processed to yield ubiquitin, reflecting the third segment that shows a much slower turnover. C) Two further proteins were tested, CTSD - Cathepsin D – (shown in orange square in A) and RPRD1A - Regulation of nuclear pre-mRNA domain-containing protein 1A (shown in orange circle in A). A cycloheximide inhibition experiment was performed on HeLa cells to block protein synthesis and thus measure the rate of protein degradation. Both CTSD (lanes 1-6) and RPRD1A (lanes 7-12) were detected by immunoblotting using specific antibodies. The western blot for CTSD shows a band at the predicted molecular weight of 44.55kDa (white arrow). However another band is also visible, slightly higher (black arrow), which shows a faster turnover across the 5 timepoints (lanes 2-6). This correlates with the average turnover of the whole protein (9.45 hours), which is much faster compared to peptides found in segment 1 from CTSD1 (26.20 hours). In relation to the second protein tested, RPRD1A, the segmentation method of analysis showed segment 1 (28.73 hours) had a significantly different turnover compared with segments 2 (7.63 hours) and 3 (14.04 hours), indicating potential isoforms with different turnover. The western blot for RPRD1A shows two bands, which correlate with the expected molecular weight of the known isoforms of RPRD1A (isoform 1 at 35.72kDa, isoform 2 at 32.92kDa and 31.63kDa). Furthermore, the upper band (black arrow, isoform 1) shows slower degradation over the timecourse (lanes 8-12) (consistent with segment 1) compared with the lower band (white arrow) (isoforms 2 and 3, consistent with segments 2 and 3).

The validity of the rule of thirds approach was confirmed by its unbiased identification of RPS27A as one of the proteins with a segment showing differential turnover to the mean value for the whole protein (see Figure 49B). In this case the mean turnover value of peptides from the carboxy terminal segment (green, ~31 hours), was ~three fold higher than the mean turnover values for the peptides in either of the other two segments (blue ~11 hours & red ~14 hours) and ~200% higher than the mean of all the peptides in this protein (~15 hours). Interestingly, the full length RPS27A protein is

expressed as a precursor that is subsequently processed to yield ubiquitin, which accounts for approximately 70% of the sequence, and a carboxy terminal segment of ~30% that corresponds to the mature ribosomal small subunit protein S27A (Chan et al., 1995). As ubiquitin is subsequently conjugated to proteins as a post-translational modification that can promote proteasome-mediated degradation, while ribosomal proteins are typically stable after incorporation into ribosome subunits, it is not surprising that these two products of the original RPS27A polypeptide exhibit different turnover values.

Two other examples were selected from the group of highlighted proteins for further analysis, corresponding to Cathepsin D (CTSD) and Regulation of nuclear pre-mRNA domain-containing protein 1A (RPRD1A) (see Figure 49C). A cycloheximide inhibition experiment was performed on HeLa cells to block protein synthesis and thus measure the rate of protein degradation. Both CTSD (lanes 1-6) and RPRD1A (lanes 7-12) were detected by immunoblotting, using specific antibodies generated by the Human Protein Atlas Project. In both cases the blotting experiments reveal two bands for each protein that decay at different rates following cycloheximide treatment (see Figure 49C, arrows). These data support the prediction from the rule of thirds analysis that the CTSD and RPRD1A proteins are expressed as distinct polypeptides with different turnover values.

7.4.3 Protein Isoform Analysis: Three in a Row Approach

A limitation with the rule of thirds approach is that not all isoforms will have structures that are separable based on analysis of arbitrary equal third regions of the protein. The available peptide coverage is also often not evenly distributed between each of these three equal segments. To provide a more general approach for predicting isoform expression, based on local clusters of peptide values, the project turned to a ‘three in a row’ method. Here, mean turnover values were calculated for each set of three consecutive peptides within the total set of peptides assigned to a given protein, moving along one peptide at a time from the amino to carboxyl terminus (see Figure 50A). The resulting mean turnover values for every group of three consecutive peptides were then plotted on the y axis, against the corresponding mean turnover value on the x axis calculated using all peptides mapped to each protein (see Figure 50B). In this plot each triple peptide mean value is shown either in light blue (default),

or in dark blue if two conditions are met. Thus, dark blue indicates that both the turnover value for that group of three consecutive peptides differs by 20% or more, (either higher or lower), than the mean value of all of the peptides assigned to that protein and that all three peptides in the group have similar values, i.e. all three are either higher, or lower, than the protein mean.

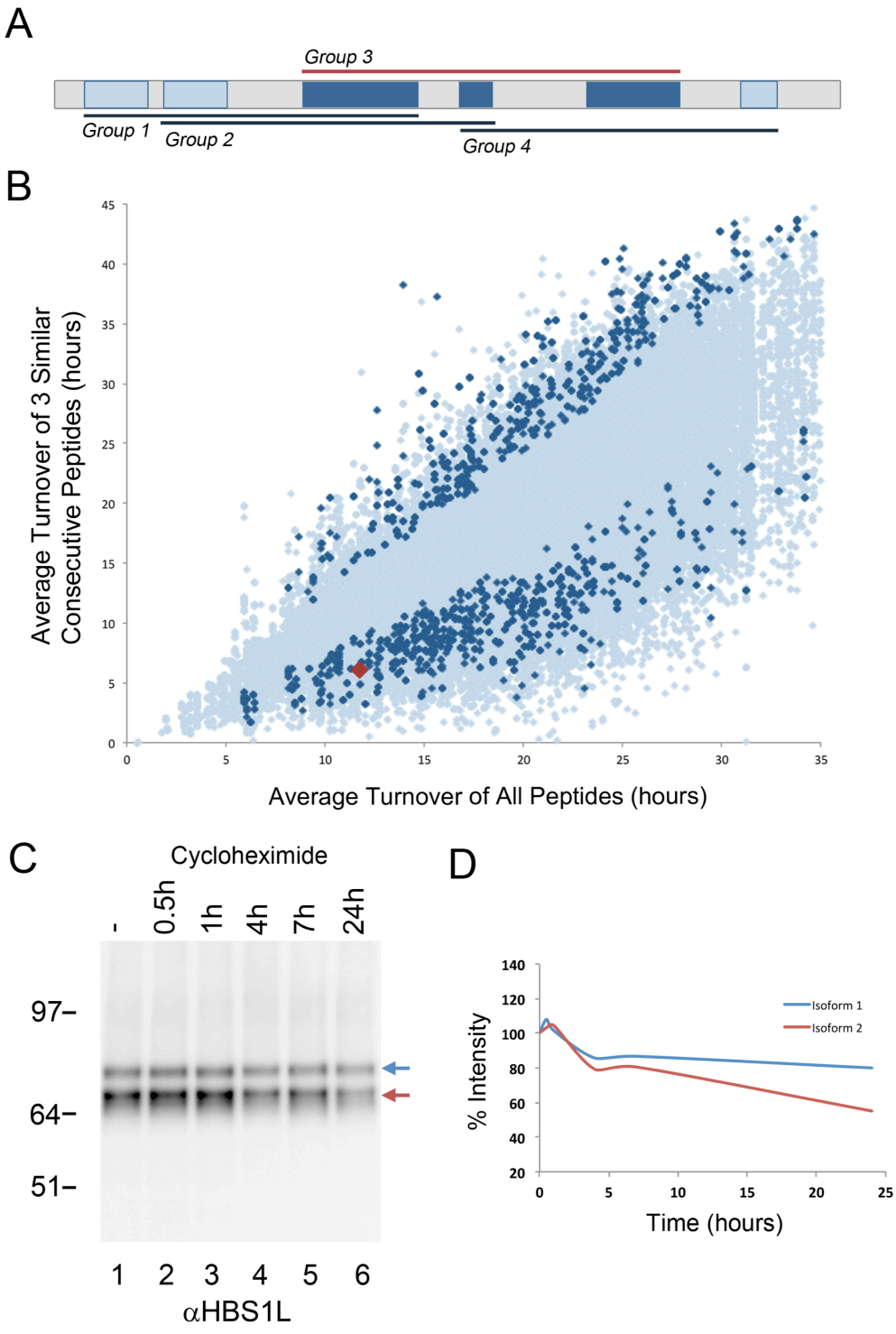


Figure 50: Protein isoform identification from consecutive peptide analysis.

In order to identify potential protein isoforms, an unbiased approach was employed whereby any region of a protein showing a very different turnover from the whole protein was highlighted. A) In order to do this, groups of three consecutive peptides were identified and the average turnover value of each group calculated. A linear representation of a protein is shown, with highlighted regions indicating identified peptides. Any group of three peptides where each peptide showed a 20% variance in turnover from the average protein turnover (calculated from all peptides) was labelled as interesting (Group 3). B) Graph showing average protein turnover from all peptides (X axis) versus the average turnover of three consecutive peptides from the protein (Y-axis). The data points highlighted in blue indicate three consecutive peptides that all have a turnover that varies by 20% greater or less than the average protein turnover calculated from all peptides, indicating potential isoforms. Highlighted in red, is protein HBS1L, HBS1-like protein, which was investigated further. C) A cycloheximide experiment was carried out to independently measure the degradation rate of HBS1L. An antibody specific for HBS1L detected two bands on an immunoblot, consistent with expression of two isoforms (blue and red arrows). These bands correlate to the known isoforms of HBS1L (isoform 1 at 75.5kDa (blue arrow), isoforms 2 and 3 at 70.13kDa and 70.63kDa (red arrow)). D) Quantitation of the two bands at multiple time points from 0.5-24 hours (lanes 2-6) following cycloheximide treatment is shown on the graph. The percentage intensity is plotted on the Y axis, across the timecourse on the X axis for the two bands found on the immunoblot. The graph shows that the two putative isoforms of HBS1L differ in their degradation rates.

For the whole cell protein turnover dataset, analysis of 178,509 groups of three consecutive peptides identified 1,790 groups (~1 %) that met this criteria and hence are shaded dark blue (Figure 50B). To validate this approach, one of the highlighted proteins was selected for which specific antibodies were available, i.e. HBS1L (red diamond in Figure 50B). A cycloheximide experiment was carried out to measure independently the degradation rate of HBS1L (see Figure 50 C & D). An antibody from the Human Protein Atlas Project specific for HBS1L detected two bands on an immunoblot, consistent with expression of two isoforms (see Figure 50C). Quantitation of the two bands at multiple time points from 0.5-24 hours following cycloheximide treatment showed that the two putative isoforms of HBS1L differed in their degradation rates (see Figure 50D). It was concluded that the three in a row approach can help to detect proteins expressed as isoforms with differential properties.

7.4.4 Isoform Analysis by Combined Protein Fractionation and Peptide MS

Protein isoforms that differ in size can be separated by chromatography prior to enzyme cleavage and MS identification of peptides. Therefore information derived from fractionation of HeLa cell proteins by 1-D SDS PAGE has been incorporated into the analysis (see Figure 51). HeLa cell extracts were separated on a 4-12% SDS PAGE

gel, which was then cut into 16 slices, numbered from the top (slice 1, largest proteins) to the bottom (slice 16, smallest proteins) of the gel (see Figure 51A). Proteins in each gel slice were digested with trypsin and the resulting peptides eluted and analysed by MS (Boisvert et al., 2011), with the resulting data plotted on a graph showing gel slice on the y axis and log predicted molecular weight of each identified protein, based on genome sequence annotation, on the x axis (see Figure 51B). These empirical data demonstrate that, as expected, the position of protein migration on SDS PAGE is positively correlated with predicted molecular weight, (Pearson correlation coefficient 0.73). In this gel system, that correlation holds true at least within the size range from ~10k-180kDa. Using the MS identification information the approximate size range of proteins migrating in each gel slice can thus be estimated. Based upon a best linear fit within the 10k-180kDa size range, the majority of proteins (~78%), migrate at their predicted molecular weight +/-40% (see Figure 51B, blue dots). Nonetheless, a substantial number of proteins identified by MS (>20%), migrate anomalously with respect to predicted molecular weight (see Figure 51B, red dots). Reasons for apparently anomalous migration is likely to include the expression of novel protein isoforms and processed polypeptides, as well as effects of post-translational modifications on migration behaviour.

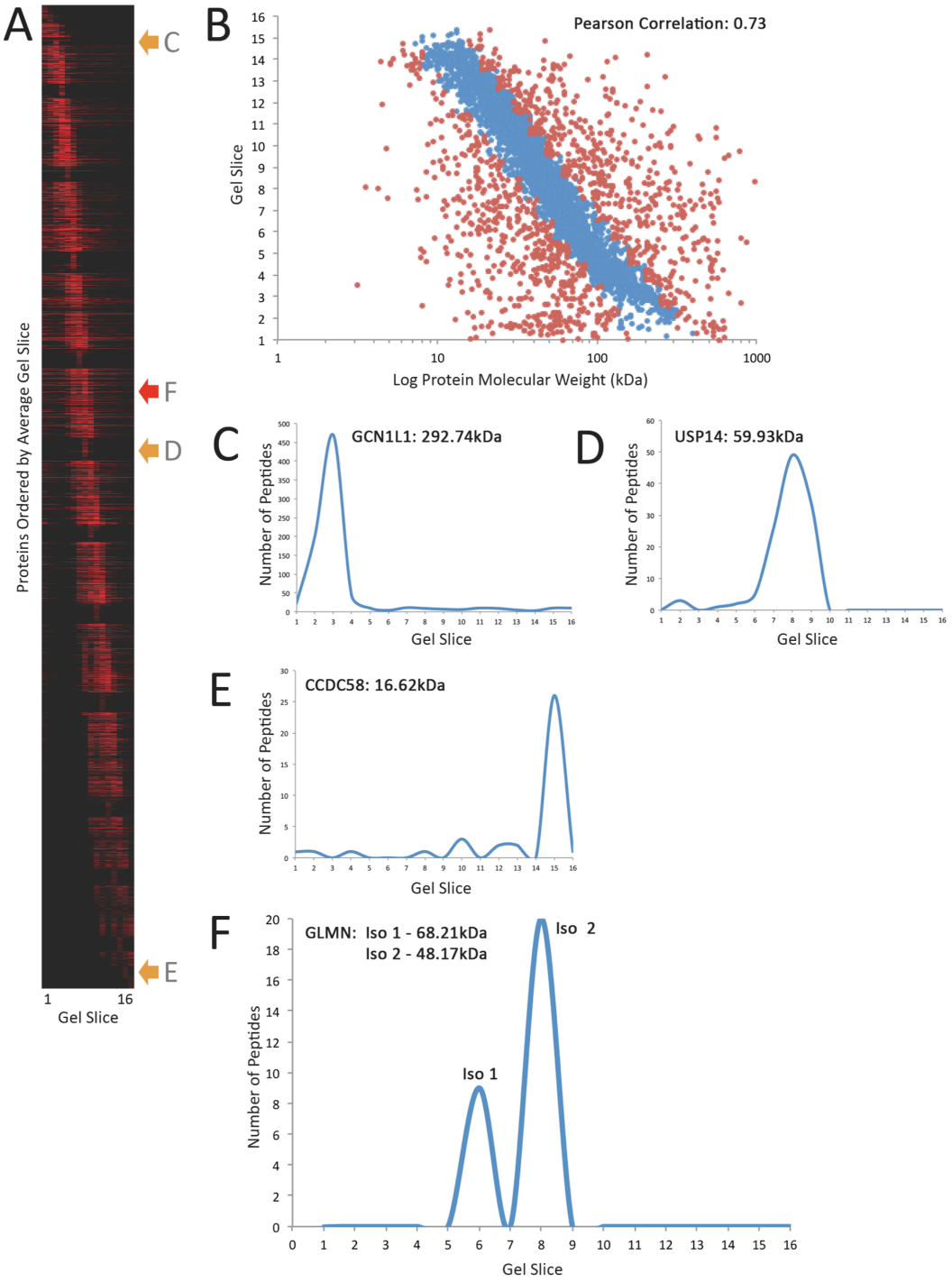


Figure 51: Protein migration study on gel fractionation.

To identify potential isoforms, data collected on 1-D SDS PAGE gel fractions was analysed. Proteins with isoforms are likely to have a difference in molecular weight and hence may migrate on the gel at different heights, hence appearing on different gel slices. A) A heat map of the 16 gel slices (horizontally) is shown with every protein identified (vertically) ordered by the average gel slice the protein was found in. This heat map shows that the proteins migrate across the 16 gel slices. B) Graph showing predicted log protein molecular weight (X axis) against gel slice (Y axis), indicating that proteins predictably migrated across the gel based on their molecular weights (Pearson Correlation: 0.73), i.e. lower molecular weight proteins at higher bands compared with high molecular weight proteins at lower bands. The majority of proteins (~77%),

migrate at their predicted molecular weight +/-40% (blue dots), however substantial number (>20%), migrate anomalously with respect to predicted molecular weight. Graphs C, D and E highlight three examples of proteins, GCN1L1 (Translational activator GCN1), USP14 (Ubiquitin carboxyl-terminal hydrolase 14) and CCDC58 (Coiled-coil domain-containing protein 58) respectively, where the peptide count (Y axis) is plotted against gel slice (X axis). These graphs show that the proteins migrate at different gel slices consistent with their molecular weight. F) Graph showing peptide count (Y axis) versus gel slice and molecular weight (X axis) for protein GLMN, Glomulin. The gel fractionation data indicates that this protein migrates at two gel slices, 6 and 8, potentially indicating the presence of isoforms. GLMN, has in fact two known isoforms, isoform 1 at 68.21kDa and isoform 2 at 48.17kDa.

Examination of the number of unique peptide identifications assigned to a given protein in each gel slice reveals the migration profile of that protein in SDS PAGE (Figure 51, C-E). For representative large (see Figure 51C, GCN1L1, 293kDa), medium (see Figure 51D, USP14, 60kDa) and small (see Figure 51E, CCDC58, 17kDa) proteins, the number of unique peptides identified shows a clear single peak across the respective gel slices. The breadth of the unique peptide abundance peak is positively correlated with protein abundance (see Figure 52), such that the most abundant proteins show broad horizontal lines in the heat map (see Figure 51A). The unique peptide count per gel slice also helps to identify distinct protein isoforms. As shown for the protein Glomulin (GLMN), which has two known isoforms of 48kDa and 68kDa, respectively. Two peaks of unique GLMN peptides are detected, centred on different gel slices (see Figure 51F). Thus, combined protein chromatography on SDS PAGE, together with peptide MS analysis, can detect the presence of protein isoforms and together with ion intensity values provides information concerning protein expression levels. Importantly, this approach can aid detection of previously unknown isoforms and/or processed and modified pools of proteins, which may have different biological properties, without prior knowledge of isoform-specific peptides or the availability of specific antibodies.

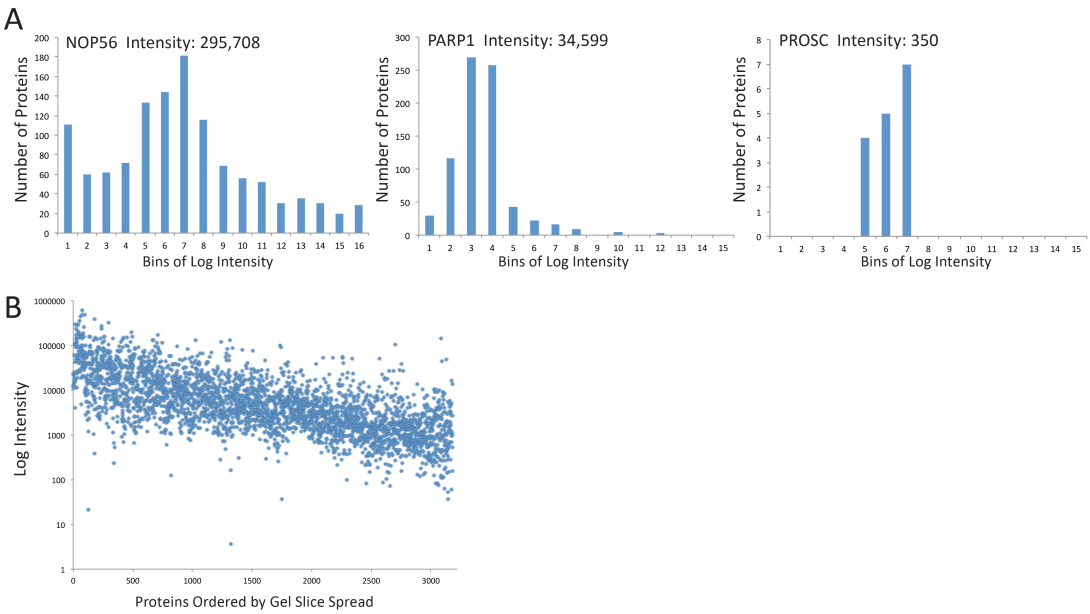


Figure 52: Protein intensity relation to gel slice

A) Graphs show number of proteins (Y axis) identified in each gel slice (X axis) for three proteins: NOP56 (Nucleolar protein 56), PARP1 (Poly [ADP-ribose] polymerase 1) and PROSC (Proline synthetase co-transcribed bacterial homolog protein). The breadth of the unique peptide abundance peak on each graph is positively correlated with the abundance of each protein. B) Graph showing Log Intensity (Y axis) against all proteins ordered by the number of gel slices the protein is found in (X axis). This graph shows that the proteins that spread across gel slices have a higher intensity compared to proteins primarily found in only one gel band.

Next, correlation analyses were performed to examine potential differences in subcellular localisation and protein turnover properties for examples of protein isoforms predicted from the combined SDS PAGE and peptide MS data (see Figure 53).

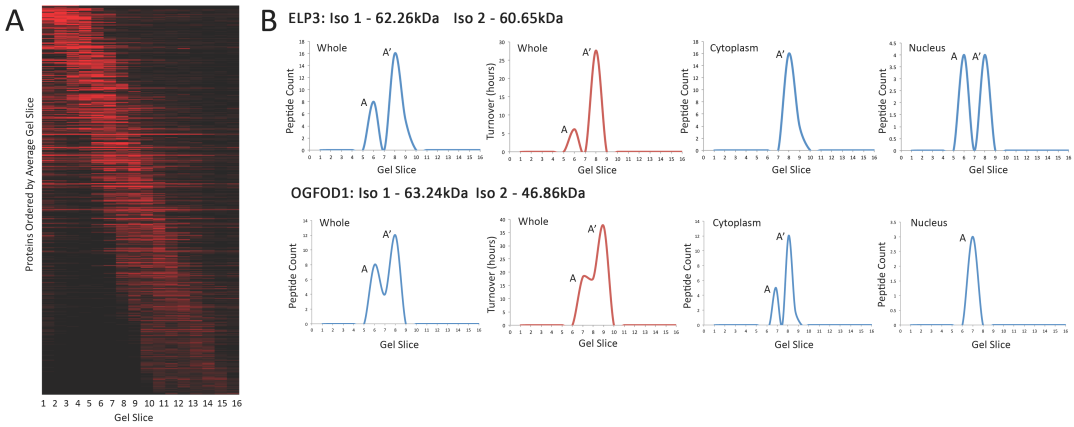


Figure 53: Protein isoform identification from gel fractionation

To identify potential isoforms from gel fractionation data, the peptide count across gel slices was analysed to identify proteins that seem to migrate at multiple gel slices. A) Heat map showing the 16 gel slices and their corresponding molecular weights (horizontally) for every protein identified (vertically) ordered by the average gel slice the protein was found in. This heat map was filtered to show only those proteins that seem to migrate at multiple gel slices. B) Two example proteins are shown, ELP3 (Elongator complex protein 3) and OGFOD1 (2-oxoglutarate and iron-dependent oxygenase domain-containing protein 1). The graphs on the left show peptide count (Y axis) versus gel slice (X axis) as an aggregate for the whole cell, indicating that both ELP3 and OGFOD1 migrate at two separate gel slices, indicating two isoforms. The cytoplasmic graph (top-middle-right) and nuclear graph (top-right) for ELP3 indicate that only one isoform is present in the Cytoplasm (A'), whereas both isoforms are detected in the Nucleus (A and A'). The turnover graph (top-middle-left), showing the turnover values detected (Y axis) in each gel slice (X axis), indicates that both isoforms of ELP3 have a different turnover, i.e. 6 hours (A) and 28 hours (A') respectively. In relation to OGFOD, the cytoplasmic graph (bottom-middle-right) and nuclear graph (bottom-right) show the isoform A is found in both the cytoplasm and nucleus, however isoform A' is only found in the cytoplasm. The turnover graph (bottom-middle-left), showing the turnover values detected (Y axis) in each gel slice (X axis), indicates that the two forms of the OGFOD1 protein at the different gel slices have different turnovers, i.e. 18.18 hours (A) and 37.49 hours (A').

By independently evaluating the SILAC data reflecting subcellular protein localisation and turnover (Boisvert et al., 2011) for the separate sets of unique peptides found in different gel slices, it can thus be predicted whether the different protein isoforms/processed forms differ in their properties. This is illustrated for proteins Elongator complex protein 3 (ELP3) and 2-oxoglutarate and iron-dependent oxygenase domain-containing protein 1 (OGFOD1), both of which are detected in two peaks of unique peptide abundance in SDS PAGE (see Figure 53B). In the case of ELP3, the larger (A) isoform has a turnover value of ~5 hours and is detected specifically in the nucleus. In contrast, the smaller (A') isoform has an apparent turnover more than five fold

slower (~27 hours) and is detected equally in the nucleus and cytoplasm. In the case of protein OGFOD1, the two isoforms detected also differ in both turnover and in subcellular distribution. The larger (A) OGFOD1 isoform has a ~50% faster turnover than the smaller (A') isoform, (~18 hours and ~37 hours, respectively). The two isoforms are differentially distributed, with the larger A isoform detected in both the cytoplasm and nucleus, and the smaller A' isoform concentrated specifically in the cytoplasm. It was concluded that this pre-chromatography approach can reveal the presence of protein isoforms with differential properties.

7.4.5 Correlating Post-Translational Modification with Protein Properties

Finally, the potential relationship between post-translational modifications and the properties of subcellular localisation, turnover and abundance measured for HeLa proteins using SILAC was investigated. In this study the effect of phosphorylation was analysed on either serine, threonine or tyrosine residues on rates of protein turnover in each of the cytoplasmic, nuclear, and nucleolar compartments (see Figure 55). Phosphopeptides were detected and quantitated for the HeLa protein localisation and turnover SILAC data set using MaxQuant. Overall, 2,444 phosphopeptides were detected and quantitated in this analysis, identifying phosphorylated residues in ~46% of the HeLa proteins (see Figure 54).

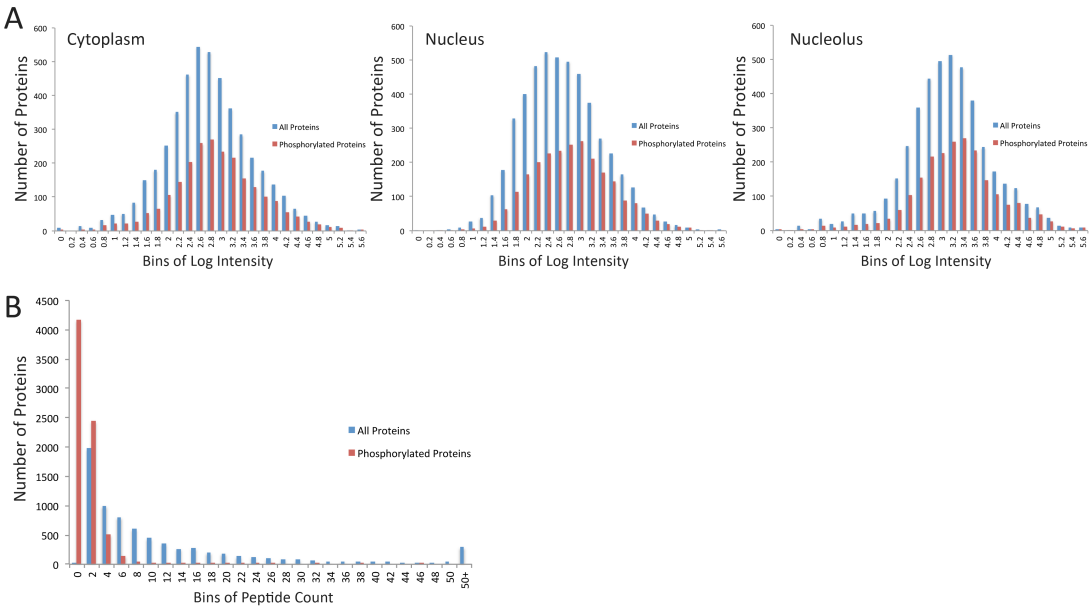


Figure 54: Phosphorylated proteins correlated with protein properties.

A) Graphs show bins of log intensity (X axis) against the number of proteins found in each bin (X axis), for cytoplasm, nucleus and nucleolus. The graphs show that the

comparison of protein abundance levels with the detection of phosphorylated peptides indicates only a weak positive correlation. This shows that the phosphopeptides studied are representative of the proteome and not reflecting the properties of only the most abundant proteins. B) Graph showing bins of peptide count (X axis) against number of proteins found in each bin (Y axis). The majority (53%) of phosphoproteins were identified with a single phosphorylated residue, although 23% had two phosphorylated peptides and 24% had three or more.

A comparison of protein abundance levels with the detection of phosphorylated peptides showed only a weak positive correlation. This indicates that the phosphopeptides studied are representative of the proteome and not reflecting the properties of only the most abundant proteins. The majority (53%) of phosphoproteins were identified with a single phosphorylated residue, although 23% had two phosphorylated peptides and 24% had three or more (see Figure 54).

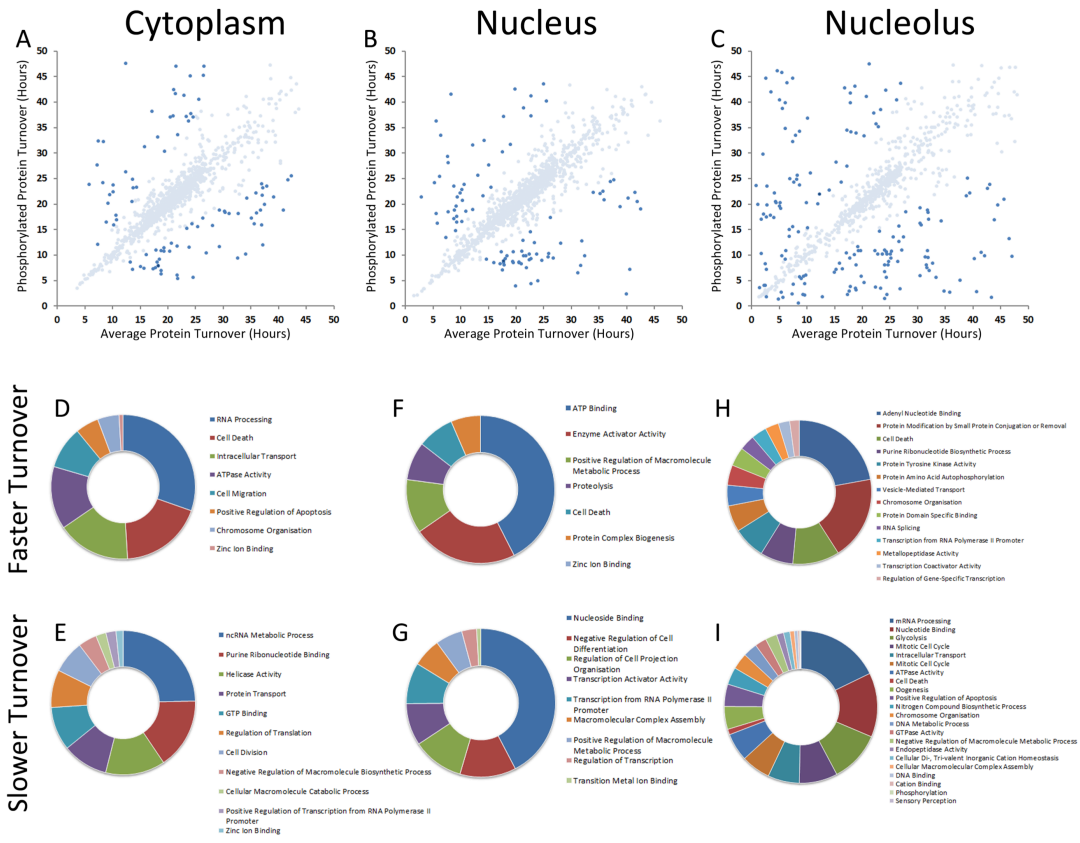


Figure 55: Phosphorylated post translation modification analysis with turnover.

Graphs A, B and C show the average protein turnover using non-phosphorylated peptides (X axis) against average protein turnover using both phosphorylated and non-phosphorylated peptides in each of the cytoplasmic, nuclear, and nucleolar compartments. Highlighted in blue are phosphorylated proteins that show a 1.5 fold change compared to the non-phosphorylated form of the protein. Comparison of graphs A (cytoplasm), B (nucleus) and C (nucleolus) show that the nucleolus has the

greatest number of phosphorylated proteins compared with the cytoplasm and nucleus. The pie charts D, E and F show the gene ontology analysis of the phosphorylate proteins that have a slower turnover in comparison with phosphorylated from and, similarly, pie charts G, H and I show the gene ontology analysis of the phosphorylate proteins that have a faster turnover in comparison with phosphorylated form.

For proteins identified in the respective cytoplasmic, nuclear and nucleolar fractions, the mean turnover value for all proteins assigned to each protein, including all phosphopeptides detected, was plotted on the y axis against the corresponding mean turnover value for all peptides assigned to the same protein, but excluding phosphopeptides, plotted on the x axis (see Figure 55, A-C). In the graphs any protein where the presence of phosphopeptides either increases, or decreases, the mean turnover value by 1.5 fold, or greater, is coloured dark blue. The data show that for most HeLa proteins the presence of one or more phosphorylated residues has little or no effect on mean turnover rate. However, a subset of proteins showed changes in turnover rate when phosphopeptides are present. Interestingly, a larger fraction of nucleolar proteins showed effects of phosphorylation on turnover rates (see Figure 55C), as opposed to either cytoplasmic, or nuclear proteins (see Figure 55 A & B). This is not caused by the subset of highest abundance nucleolar proteins, such as ribosomal proteins and nucleophosmin, suggesting that there is a broader effect of phosphorylation on modulating nucleolar protein turnover rates.

Gene ontology analysis was carried out to categorise the phosphorylated proteins showing the greatest increase (see Figure 55 D, F & H) and greatest decrease in turnover (see Figure 55 E, G & I), for the cytoplasmic (see Figure 55 D & E), nuclear (see Figure 55 F & G) and nucleolar (see Figure 55 H & I) compartments, respectively. This shows specific groups of proteins whose turnover rates are most affected by phosphorylation. This includes ATP and nucleotide binding proteins, multiple cell cycle regulated proteins and proteins involved in apoptosis and cell death response mechanisms.

7.5 Discussion

This study has investigated multiple data analysis approaches that can be used to identify the expression of protein isoforms that exhibit differential localisation and/or turnover properties. Examples of protein phosphorylation correlating with altered turnover rates were identified in different subcellular compartments. These analyses

are performed on SILAC-based quantitative mass spectrometry data from fractionated HeLa cells, where changes in isotope ratios are used to measure turnover rates in the separate cytoplasmic, nuclear and nucleolar compartments (Boisvert et al., 2011). It has been shown that separating intact proteins by chromatography, prior to enzyme digestion and peptide identification by mass spectrometry, can be effectively combined with SILAC analysis of changes in peptide isotope ratios to identify protein isoforms and assess whether the isoforms have different properties. Collectively, these experimental procedures and data analysis approaches provide a new framework for the systematic detection and analysis of protein isoforms and their biological properties that can be used to expand the functional annotation of the genome.

Differential proteomic analysis using SILAC involves measuring differences in the ratio of separate isotopic forms of the same peptide, which in turn is related to a specific biological property according to the experimental design. Thus, differences in isotope ratios can be used, inter alia, to measure changes in protein expression levels following drug treatment, to discriminate between specific and non-specific protein interaction partners or to compare subcellular protein localisation. Typically, mean values are calculated for the different isotopic forms of all of the peptides detected that map to a given protein, as deduced from genomic sequence information. A potential limitation with this strategy however is that it usually does not discriminate between peptides arising from functionally distinct pools of protein encoded by the same gene. Thus, ensemble measurements are generated that can average the separate properties, or responses, of two or more distinct protein isoforms. It has been shown here that, at least in part, it is possible to circumvent these limitations and to identify protein isoforms and compare their properties, both using information provided by detailed analysis of isotope ratios for separate peptides assigned to the same protein group and by incorporating information from protein chromatography prior to enzyme digestion and MS analysis.

The candidate peptide approach is conceptually simple and can be effective, as demonstrated here for the protein NudCD1 (see Figure 48). However, it is often not possible either to identify, or to reliably detect and quantitate, isoform-specific peptides. This restricts the use of the candidate peptide approach to the analysis of protein isoforms whose structures are already characterised and where one or more

isoform-specific peptides have been identified. Even in these cases, quantitation of the isoform response is often derived from analysis of only one or two specific peptides, which can reduce the overall reliability of the measurements. It has been shown that more promising general approaches for detecting isoforms and comparing their properties involve the systematic evaluation of mean isotope ratio values for groups of peptides within the total set of peptides mapped to a specific gene. Importantly, with both the 'rule of thirds' and 'three peptides in a row' approaches, analysis of the SILAC data can predict the potential existence of either isoforms, or processed forms of proteins, as well as compare their properties, without prior knowledge of either isoform structures, or expression. In each case, the mean isotope ratio values of sub-groups of peptides can be evaluated with respect to the mean value, either for all the peptides in the protein, or for values for neighbouring groups of peptides, or both. Objective statistical criteria can be applied to these comparisons that will aid the reliable detection of isoforms and thereby help to annotate the functional expression of the genome.

This study has validated the effectiveness of data analysis strategies involving statistical comparisons of isotope ratio values for local clusters of peptides within a protein. Several ways are envisioned in which such general approaches can be enhanced further in future. For example, using filters that compare more closely variations in values between peptides in a group and by defining peptide groups with reference to 3-D crystal structure information on proteins. Whatever future refinements are made to the data analysis procedures, it is clear that a critical point is having a high quality SILAC data set for the proteome under study and in particular having as wide a peptide coverage as possible for each protein. The HeLa SILAC data set studied here included over 80,000 peptides from ~8,000 proteins, with an average coverage of ~10 peptides per protein identified. Recent analyses indicate that this is already a large enough sample of the expressed HeLa proteome to be highly representative of the general behaviour of cell proteins (Boisvert et al., 2010). In future, therefore, the project will aim to expand the number of peptides analysed, not primarily to increase the total number of proteins identified, but rather seeking to enhance the peptide coverage for each protein. It is anticipated this will aid the unbiased detection of protein isoforms and their properties that can in turn be related to biological mechanisms and responses.

In most cases, differences in structure between protein isoforms alters their size and/or charge, which in turn provides an opportunity to separate them by chromatography, as demonstrated here using 1-D SDS PAGE. The results show that independently evaluating the differences in peptide isotope ratios for the same peptides migrating in different chromatographic fractions (in the present case different gel slices), provides a powerful approach for detecting protein isoforms and assessing differences in their properties. Combining fractionation of protein extracts with downstream enzyme cleavage and MS analysis thus provides important information that is lost in procedures where entire extracts are digested without pre-fractionation and peptides analysed en masse. The isoform information is similarly lost if extract fractionation is performed at the peptide, rather than protein level. To provide higher resolution separation of isoforms, therefore, it is planned in future to increase the degree of protein fractionation prior to MS analysis. For example, using 2-D fractionation of extracts, combining ion exchange and gel filtration chromatography. It is anticipated that such 2-D protein fractionation strategies, combined with increased peptide coverage, will further enhance the efficiency of detecting isoforms and characterising their properties.

It has been shown previously that the subcellular distribution of the proteome can be measured using a SILAC strategy where different cell compartments and organelles are isotope-encoded (Boisvert and Lamond, 2010, Boisvert et al., 2010, Boisvert et al., 2011). It has been shown also that system-wide changes in protein localisation could be measured in response to drug treatment and in cells with different genotypes. Here the ‘spatial proteomics’ approach has been extended to detect protein isoforms that are differentially localised within the cell and to analyse differential effects of protein phosphorylation on turnover in different subcellular compartments. This can be developed further in future in several ways. First, a higher resolution map of proteome localisation can be derived by more extensive cell fractionation prior to protein chromatography and MS analysis. For example, the cytoplasmic compartment can be sub-fractionated into plasma membrane, cytosol and organelle fractions and work is underway to implement this. Second, many other post-translational modifications in addition to phosphorylation can be analysed and their potential effects on the properties of specific protein families and protein isoforms evaluated and compared in different cellular compartments. Third, the analyses to date have analysed mixtures of

cells at different cell cycle stages. However, it is already known for specific proteins that their expression levels and properties, including localisation and PTMs, can change during different stages of interphase and mitosis. It is therefore planned to expand future studies to encompass system-wide, quantitative analysis of the properties of protein isoforms both in multiple subcellular locations and at different cell cycle stages. The resulting data are likely to provide a major source of information that can reveal unexpected and novel molecular relationships and potential regulatory mechanisms for future investigation.

7.6 Distribution of Effort

Lamond Laboratory biologists, primarily Francois-Michel Boisvert, carried out the experimental bench work during this study. The analysis described in this chapter was a joint effort between Yasmeen Ahmad and Francois-Michel Boisvert.

Chapter 8: Discussion

The Human Genome Project was a major international endeavour aimed at identifying and mapping all genes, which control hereditary characteristics in living organisms. A gene is any given segment along a DNA strand that encodes instructions allowing a cell to produce a specific product - typically a protein. However, since the completion of the human genome project (2003) the focus of research has changed from working at the genome level, identifying and mapping genes, to documenting the function of genes and realising how changes in the sequence relate to health and disease at the cellular level. By researching the proteins expressed by genes under various conditions, the field of life sciences has made significant contributions to the understanding of how the human body functions. This research has led to the definition of a new field: proteomics, which aims to discover, annotate and describe the properties of proteins in living organisms.

The Lamond Laboratory, based in the Wellcome Trust Centre for Gene Regulation & Expression at the University of Dundee, is playing a role in the development and application of new quantitative and high throughput methods for the analysis of gene expression and cell biology. In particular the Lamond group focus on how cancer and other diseases can result in changes in the spatial distribution, stability and function of proteins in human cell lines. These studies involve large-scale use of proteomics technologies, including novel mass spectrometry (MS) approaches based upon stable isotope labelling (i.e. SILAC). A key feature of these quantitative MS-based methods is the creation of huge volumes of data that are impossible to analyse by manual inspection. These data are major resources that require new approaches to manage, analyse and store efficiently. These problems are further enhanced by the complexity of the data, the non-consensus on data formats and non-existent data standards. Furthermore, there is a need to archive these data in accessible repositories to promote sharing of data. Hence, it is imperative that these advancements in science are supported by adequate developments in the field of computing. There are many new and exciting discoveries to be made through cross discipline work that brings together life sciences research and computing in a usable fashion, which can enhance knowledge and understanding in both fields.

"The mapping of complex proteomics data to biological processes has become impossible by manual means, and the need for computer-aided data analysis is essential for further progress in the field." (Kumar and Mann, 2009)

To date there is no method to routinely capture, manage and archive datasets from such studies. Furthermore, downstream analysis is a major challenge posed by proteomics technologies. Advances in technology have allowed for more sophisticated proteomics experiments, which have resulted in generation of an increased volume and complexity of data that demands the development of new tools due to the inadequacy of existing software, such as Excel. In these situations, biological researchers are often forced to carry out minimal analysis manually and then hand-over their datasets to bioinformaticians who have the necessary computing skills to handle these data. This is frustrating as the biologists are experts in how the data are generated and having driven the formulation of the initial hypothesis that led to the experiments and data generation, they are more acutely aware of how they would like to question the data further and its potential. Furthermore, unless there is extensive interaction between bioinformaticians and biologists, there will be in minimal information exchange regarding the context of the data and the processes involved in generating the data.

"Despite our reliance on computation, most scientists are not capable of complex data storage and analysis computing, and therefore rely on computer programmers to do this for us." (Proteomics Researcher, Lamond Laboratory)

During this thesis a consolidated data environment has been created that provides a central source for project-wide decision-making. The challenges described above are met through a pipeline for quantitative data derived from proteomics experiments. This pipeline has been incorporated into the development of a software suite: PepTracker (http://www.peptracker.com), which provides a Laboratory Information Management System (LIMS) and supports the upload of datasets processed by MaxQuant (current, future and legacy versions of 3rd party software). PepTracker incorporates an easily accessible data repository that allows sharing of the uploaded data, visualisation interfaces to navigate the datasets, tools to assess quality of constituent data, interactive graphical control of basic analysis and more advanced features for global analysis.

Benefits of having a repository include access to a large collection of baseline datasets that are labelled with detailed metadata. These datasets can be analysed together to aid in the analysis of new datasets, through confirmation of trends and identification of false-positive results. As the experimental protocols and analysis become more complex, manual analysis of single datasets and the resultant observed patterns give rise to subjective errors. In addition, depending on the experiment and the protocols used to create samples, contamination may creep into samples from the procedures carried out. However, for accurate analysis MS datasets must be of high confidence and users must be sure that data repositories sharing such data contain similar high-accuracy and high-confidence datasets. Through automated approaches, software can evaluate and normalise multiple datasets to tackle these concerns and then allow for more reliable automated classification, visualisation and clustering of datasets, leading to biologically interpretable results and insights. The work in this thesis addresses these issues and enables new discoveries through entirely novel analysis strategies which take full benefit of the uniquely quantitative nature of the proteomics data being generated.

Using this repository of datasets in PepTracker, this thesis led to the development of a specific analysis tool, the Protein Frequency Library (PFL), in order to tackle the issue of contamination within samples (Boulon et al., 2010a). Using the multi-dimensional datasets allowed the research to explore the use of techniques from other fields, including Business Intelligence (BI), to tackle the analysis of data from a global data standpoint in contrast to the typical one user - single experiment analysis approach common in life sciences.

The use of BI is not previously documented in biology or proteomics and is rarely found in research science in general. With the global BI approach, each dataset can enhance the analysis of every other dataset in the repository, allowing researchers to make better, informed decisions that result in improved, more efficient science (in terms of human and physical resources). This type of analysis has been coined as “super-experiment” analysis within this thesis, whereby datasets are analysed collectively rather than individually. The whole in this case is much more than the sum of the individual datasets and can answer questions that were not conceived of when the original experiments were performed. A key feature of the super experiment

concept is that each new dataset that is added improves the analysis of all future experiments and also allows re-analysis of previous data to detect trends and relationships not apparent when prior experiments were first performed.

The application of business intelligence to proteomics in this research has facilitated reliable identification of protein interaction partners (see Chapter 5: Multidimensional Analysis with IP Experiments). Techniques from the BI field were used to perform multidimensional data analysis, in order to improve the discrimination between specific and non-specific protein associations and to analyse dynamic protein complexes. These strategies involved annotating the frequency of detection in immunoprecipitation experiments for all proteins in the human proteome. From this annotation, the likely specific interaction partners could be discerned more reliably as these were usually expected to have a low frequency. This list of proteins and annotation produced a Protein Frequency Library (PFL) that improves on previous use of static “bead proteomes”. The PFL produced not only provides a flexible and objective filter for discriminating between contaminants and specifically bound proteins but it can be used to normalise data values and facilitate comparisons between data obtained in separate experiments. The PFL is a dynamic tool that can be filtered by specific experimental parameters to generate a customised library. Furthermore, it is continuously updated as data from each new experiment are added to PepTracker thereby progressively enhancing its utility. The application of the PFL to pull-down experiments is especially helpful in identifying either lower abundance, or less tightly bound, specific components of protein complexes that are otherwise lost amongst the large, non-specific background.

The PFL has been fully implemented in a PFL Viewer tool (http://proteinfrequencylibrary.com) that provides an intuitive graphical user interface for researchers wishing to explore the PFL database, which is implemented as an OnLine Analytical Processing cube. Furthermore, the PFL functionality is built into the main PepTracker suite so that researchers can integrate the PFL with their datasets to allow normalisation using the wealth of experience captured by the PFL.

The PFL Viewer is one tool that captures the concept of super-experiments to formulate information that can be used to annotate proteins with an additional confidence score. This type of annotation is valuable to researchers as it can help

direct efforts more systematically to proteins that may yield interesting biological results. This is imperative for biologists before they follow lines of research, which may take them on a detailed and costly study of a particular protein/set of proteins. Biologists will focus a number of years on such studies, hence heightening the importance of being able to identify research worthy proteins.

Within the Lamond Laboratory, having the PepTracker system in place, researchers have been driven to carry out larger scale proteome-wide studies that can provide additional protein annotations and proteome level analysis capabilities. One of these large-scale studies has been the study of global proteome turnover rates of human proteins, conducted by Francois-Michel Boisvert and Fabien Charriere (see Chapter 6: Spatial Localisation & Turnover Analyses). This experiment collected abundance, localisation, synthesis, degradation and turnover data over time for different cellular compartments, hence making these data an extremely valuable resource, which can be mined to answer questions such as the variation in turnover trends across different cellular compartments (Boisvert et al., 2011). These data can also be of worth to other researchers wanting to gain further insight into specific proteins of interest. Hence, making these data available through an online tool would benefit many researchers. The Turnover Viewer was created as a practical utility that can be accessed by biologists globally. It documents the turnover, half-life, abundance and localisation information collected during this proteome wide study and links this information to spatial datasets to provide “super-experiment” analysis that could add further annotation to the human proteome.

Leading on from this work, new analysis approaches have also been developed with the aim of extracting more value from the data generated. Using the global proteome turnover and localisation dataset, strategies were developed to carry out more accurate protein identifications by applied knowledge analysis (see Chapter 7: Protein Isoform, Localisation and Turnover Analysis). In higher eukaryotes, many genes encode two or more protein isoforms and the properties and biological roles of these separate isoforms are often poorly characterised. In a simple analysis, these isoforms can be overlooked and misidentified. Within this research, using a multidimensional approach to the analysis brought together various dimensions in the study to accurately identify either protein isoforms or functionally distinct pools of proteins that differ in

measurable properties. A number of systematic approaches were developed and published for the detection of protein isoforms with differential biological properties (Ahmad et al., 2011). These strategies focused on identifying proteins that contained identified peptides with SILAC ratios that were off trend. Furthermore, protein isoforms were also detected by independently evaluating gel fractionation data that displayed anomalies, suggesting the presence of multiple forms of a protein. Examples of protein phosphorylation correlating with altered turnover rates were also observed in different subcellular compartments. These strategies validate the benefits of a multi-dimensional strategy to data analysis. Collectively, the experimental and analytical approaches developed here provide a framework for expanding the functional annotation of the genome.

This research has paved the way for novel methods of analysing MS data. Through the PepTracker suite the research has been able to record metadata and capture the experimental experience of researchers. The visualisation and analysis built into PepTracker allows users to navigate datasets and formulate hypotheses with ease – a task that was previously becoming impossible. By realising the potential of collecting proteomics datasets, this thesis has evidenced the validity of analysing datasets together to form new hypothesis that span the whole human proteome. Furthermore, these broad spanning analyses have provided further protein annotation to guide research activities by providing methods of accurately identifying proteins and assessing genuine interaction partners, as well as building up a dynamic picture of cellular proteomics through large-scale proteome studies and data analyses.

The Turnover and PFL Viewers are already available online for the benefit of the academic community. It is envisioned that other elements of this pipeline and software are also well established to be extended for more general use out with the Lamond Laboratory experimental protocols.

The approach to software development in this thesis has flourished due to clear understanding of the importance of communication between end users and myself, the developer. The challenges affecting a software development project in life sciences centre on domain understanding of a specialised science, knowledge extraction from expert biologists, dealing with levels of experience in computer literacy and analysis, discovery of unexpected requirements due to the pioneering nature of the field and

the risk of wrong decisions. So far I have worked whilst being embedded directly in the research laboratory, which has allowed for good dialogue throughout the software development process. By understanding the importance of end researchers, rapid deployment and an iterative process, it has been possible to make rapid progress in both the computational and biological aspects of the project.

8.1 Proteomics Data Management and Analysis in the Future

It is imperative that advances in many areas of life sciences research, especially high throughput approaches that generate large volumes of data, are supported by adequate developments in the field of computing. There are many new and exciting discoveries to be made through cross discipline work that brings together science and computing in a usable fashion that can enhance knowledge and understanding in both fields. By having an open exchange of ideas and close collaboration between cell biologists and computer scientists, exciting new solutions can be created that are researcher friendly and intuitive, whilst providing researchers with more effective representations and interpretations of their data.

Collectively, the experimental procedures and data analysis approaches described in this thesis provide a new framework for the systematic detection and analysis of proteins that can be correlated with biological properties and hence used to expand the functional annotation of the genome and take it to the next level.

Proteomics has reached a critical stage where it is now possible to produce high quality datasets with wide peptide coverage of protein identifications. This will only continue to improve over the years to come, with data growth being described as exponential. The measurement of protein properties for whole proteomes in different cell types, under different growth conditions and at many time points generates very large volumes of data. This allows proteomics a definite position in the ‘Big Data’ arena with predicted sample raw files growing towards the 10GB size in the next 2-3 years, leading to generation of terabytes if not petabytes of data. This data growth will require further sophisticated data management and pipelining strategies to prevent researchers from drowning in their data. Storage of these data is also another consideration that will become ever more prevalent. A subset of these data will be in the form of processed and annotated data outputs that are the primary resource used

by the biological community. A large amount of data will be in the form of archived raw files that must be stored securely and available for reference if called upon.

There is further development work that can be carried out in the future to extend the core functionality of the well-documented existing code base of the PepTracker suite, with the intention of expanding the software for a broader audience. This requires further attention to dealing with a larger user base, additional levels of security and a much more varied experiment set. The software has been created with the intention of scalability and in awareness of the fact that science is constantly advancing. Hence the metadata to be collected regarding experiments is continuously evolving. However, with a wider user base the experimental variety increases and hence resources could be applied to extending the software to deal with the additional requirements and features specific to other laboratories.

In addition, the current web-based software can be extended to run as a cross platform enterprise desktop application. With the increased size of datasets it is more desirable for users to have the ability to run their analysis on a desktop platform as well as a web platform. This provides increased power and capabilities for larger datasets that are more difficult to manipulate across the web. Also, with a desktop platform combined with a local database, users can download datasets for offline access. The proof of concept and suitability of the software has already been proven through the creation of the fully functional web based version of the software.

Increasingly researchers are demanding software improvements for faster analytics to open up current bottlenecks in the quantitative proteomics data pipeline. This requires that current software running on single desktops be scaled up to cluster level access and data warehousing solutions to become more prevalent. PepTracker is one of the first quantitative proteomics software to provide data warehousing capabilities. Additionally, pipelining the PepTracker workflow with commercial and/or freeware/shareware software, that allows MS identification and quantification, would provide researchers with a smoother pipeline for MS analysis.

The visualization and analytics interfaces that have been created thus far have already proven to be immensely useful. However, these can be further developed with concentrated effort on the maximisation of interaction with data. The existing

software includes charting and graphing capabilities that integrate mouse controls for zooming, scrolling and hover-over tooltips, cross comparison of datasets, management of protein groups and automated classification and filtering of contaminant proteins. By extending these existing tools, it is possible to enable more enhanced analysis of datasets. One way in which this can be implemented is via an Application Programming Interface (API) that would provide users, who are proficient at coding, with the capability of building their own plugins that can be used with the PepTracker suite. This API would provide researchers with programmatic access to their datasets, which in turn enables novel analysis possibilities. An API would also provide additional flexibility and encourage users to take part in the development efforts. Due to the varied nature of the science carried out in laboratories around the world, it is impossible to cover all potential functionality that may be requested by external researchers, hence creating an API raises the prospects of a much wider user base.

Using the PepTracker suite, the work in this thesis has already established and validated successful workflows for the large-scale, quantitative measurement of key protein properties, including subcellular protein localisation, the reliable identification of specific protein interaction partners, the systematic identification of protein isoforms and protein pools, the measurement of protein synthesis, degradation and turnover rates and the identification and correlation of patterns of post-translational protein modifications.

Increasingly the Lamond Laboratory is moving from studying static snapshots of the cell, to carrying out experiments that can aid in the development of a dynamic picture of the cell. Within a cell, the structures and properties of proteins are crucial for their function and can vary greatly. Subcellular localisation patterns, post-translational modifications, rates of synthesis and degradation and interactions with partner proteins are all variable. Furthermore, all of these properties not only vary between proteins, they are also dynamic and can vary for the same protein at different times, depending on parameters such as cell cycle progression, growth rate and signalling events. Proteomes are thus inherently complex and their properties in constant flux. This is the challenge to be explored in future proteomic studies within the Lamond Laboratory, aiming to ideally not only identify which proteins are expressed in a cell or

organelle, but also characterise their properties and quantify how these change in response to different perturbations and cell cycle stages etc.

Building on the work in this thesis, the Lamond Laboratory can move forward and extend the 'spatial proteomics' approach by compiling an even higher resolution map of proteome localisation through more extensive cell fractionation prior to protein chromatography and MS analysis. For example, the cytoplasmic compartment can be further sub-fractionated into cytosol and organelle fractions and work is underway to implement this. Furthermore, additional post-translational modifications can be added to the analysis and their potential effects on the properties of specific protein families and protein pools evaluated and compared in different cellular compartments. In addition, the analyses to date have analysed mixtures containing cells at different cell cycle stages. However, it is already known for specific proteins that their expression levels and properties, including localisation and PTMs, can change during different stages of interphase and mitosis. It is therefore important to expand future studies to encompass system-wide, quantitative analysis of the properties of proteins both in multiple subcellular locations and at different cell cycle stages. The resulting data are likely to provide a useful source of information that can reveal unexpected and novel molecular relationships and potential regulatory mechanisms for future investigation.

Over recent years, the Lamond Laboratory has shifted resources to increase analysis capabilities from studying single datasets to analysing the global proteome and system-wide changes at the cellular level. PepTracker and the idea of super experiments have thus far highlighted the huge potential for further development of these approaches in the field of proteomics. The super experiment concept is relatively unexplored and can be developed much further in the realm of biology by extending the work described whilst accessing more sophisticated analysis techniques. Through focussed application of new mass spectrometry based proteomics tools, study and documentation of dynamic protein properties on a proteome wide scale, researchers can discover and annotate proteins with information that can eventually lead to major advances in cell biology and aid the development of new approaches for the pharma and healthcare industries to evaluate drug toxicity and reduce the costs inherent in bringing safe new drugs to market.

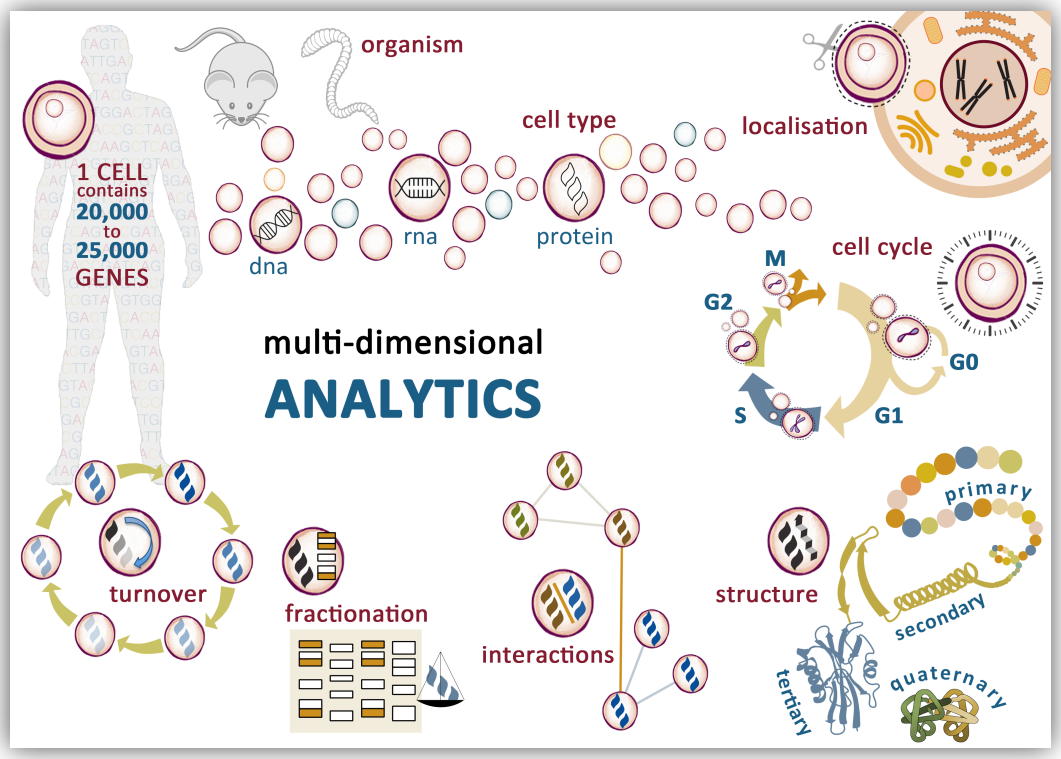


Figure 56: Multidimensional analytics of protein properties.

Due to the highly complex nature of the new types of data being gathered, there is a need for novel ways to visualise and interface with this biological data. Traditional software approaches mainly deal with visualising data using two or three dimensions at any one time. In the future, PepTracker can be expanded to explore the use of many more dimensions on the same visualisations, including shape, colour, size, labels, gradients etc. Each dimension in a dataset can have different characteristics, which defines their suitability for visualisation using various techniques. This would be particularly useful for displaying complex multidimensional datasets in a format that is easy for researchers to rapidly assimilate and interpret. Through extending multidimensional analytics it will be possible to uncover new trends and relationships within the data.

There is also a huge potential for integrating the proteomics data with other forms of biological data (e.g. sequence data and imaging data) to expand further the value of the information recorded in each proteomics experiment. Dealing with heterogeneous data will also mean developing automated quality control. It is anticipated that the techniques developed here will be researched and developed further to allow

evaluation of datasets based upon the similarities/dissimilarities found in a vast collection of historical data.

The PepTracker data warehouse is continuing to grow in the Lamond Laboratory as further datasets are added to the data warehouse, both those generated in-house and datasets obtained from collaborators. Thanks to the variety of the experiments carried out in the Lamond Laboratory, PepTracker is accumulating and recording diverse types of experimental data that could be used to test new mining techniques as they become available. This data collation can be further improved with data from external laboratories. Future directions involve supporting a larger group of researchers and sharing the resources and data outputs with the international community.

The PepTracker environment is in daily use and in active development within the Lamond Laboratory, however there is growing demand for access to PepTracker from the wider international research community, both in Dundee and globally from academia and pharma industry. Interest has been registered from institutes globally, who would like to be involved in the development work through contribution of data and/or use of our current and proposed analysis and visualisation tools. There has also been interest and dialogue with the pharma industry who have expressed their interest in making use of the proposed tools and data to make the arduous process of developing new medicines much cheaper and safer in the future. Moving forward a major aim could be to expand PepTracker to support a larger group of users and share the resources and data outputs with the international community, including commercial companies. Currently PepTracker is deigned to run locally, by developing the application further, it would be possible to make it more distributable with a ‘download and install’ version of the PepTracker software and tools, which can be made available to the academic community and licensed commercially.

Through the already implemented accurate tagging and aggregation of complex quantitative data, this thesis provides the basis for developing pioneering new approaches that allow biologists to carry out multi dimensional analysis, benefitting both basic and advanced biomedical projects around the world. PepTracker is already collecting vast amounts of data, which can be used as a test bed for mining. This thesis has highlighted the benefits of focusing on data management and visualisation tools to aid access, interpretation and manipulation of the large volumes of data being

generated by cell biologists. With the focus on enhanced user interaction with data, this thesis is aiding in the concerted, large-scale characterisation of cell proteomes by the Lamond Laboratory. The analytical techniques developed in this thesis will be transferrable to many more types of data. In the future it is possible for the PepTracker tools to be used out with research labs, in commercial drug discovery companies.

References

AEBERSOLD, R. & GOODLETT, D. R. 2001. Mass spectrometry in proteomics. *Chem Rev,* 101**,** 269-95.

AEBERSOLD, R. & MANN, M. 2003. Mass spectrometry-based proteomics. *Nature,* 422**,** 198-207.

AHMAD, Y., BOISVERT, F. M., LUNDBERG, E., UHLEN, M. & LAMOND, A. I. 2011. Systematic Analysis of protein isoforms and modifications affecting turnover and subcellular localisation. *Mol Cell Proteomics*.

ANDERSEN, J. S., LAM, Y. W., LEUNG, A. K., ONG, S. E., LYON, C. E., LAMOND, A. I. & MANN, M. 2005. Nucleolar proteome dynamics. *Nature,* 433**,** 77-83.

APWEILER, R., MARTIN, M. J., O'DONOVAN, C., MAGRANE, M., ALAM-FARUQUE, Y., ANTUNES, R., BARRELL, D., BELY, B., BINGLEY, M., BINNS, D., BOWER, L., BROWNE, P., CHAN, W. M., DIMMER, E., EBERHARDT, R., FEDOTOV, A., FOULGER, R., GARAVELLI, J., HUNTLEY, R., JACOBSEN, J., KLEEN, M., LAIHO, K., LEINONEN, R., LEGGE, D., LIN, Q., LIU, W. D., LUO, J., ORCHARD, S., PATIENT, S., POGGIOLI, D., PRUESS, M., CORBETT, M., DI MARTINO, G., DONNELLY, M., VAN RENSBURG, P., BAIROCH, A., BOUGUELERET, L., XENARIOS, I., ALTAIRAC, S., AUCHINCLOSS, A., ARGOUD-PUY, G., AXELSEN, K., BARATIN, D., BLATTER, M. C., BOECKMANN, B., BOLLEMAN, J., BOLLONDI, L., BOUTET, E., QUINTAJE, S. B., BREUZA, L., BRIDGE, A., DECASTRO, E., CIAPINA, L., CORAL, D., COUDERT, E., CUSIN, I., DELBARD, G., DOCHE, M., DORNEVIL, D., ROGGLI, P. D., DUVAUD, S., ESTREICHER, A., FAMIGLIETTI, L., FEUERMANN, M., GEHANT, S., FARRIOL-MATHIS, N., FERRO, S., GASTEIGER, E., GATEAU, A., GERRITSEN, V., GOS, A., GRUAZ-GUMOWSKI, N., HINZ, U., HULO, C., HULO, N., JAMES, J., JIMENEZ, S., JUNGO, F., KAPPLER, T., KELLER, G., LACHAIZE, C., LANE-GUERMONPREZ, L., LANGENDIJK-GENEVAUX, P., LARA, V., LEMERCIER, P., LIEBERHERR, D., LIMA, T. D., MANGOLD, V., MARTIN, X., MASSON, P., MOINAT, M., MORGAT, A., MOTTAZ, A., PAESANO, S., PEDRUZZI, I., PILBOUT, S., PILLET, V., POUX, S., POZZATO, M., REDASCHI, N., et al. 2010. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research,* 38**,** D142-D148.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics,* 25**,** 25-9.

BABU, M., KROGAN, N. J., AWREY, D. E., EMILI, A. & GREENBLATT, J. F. 2009. Systematic characterization of the protein interaction network and protein complexes in Saccharomyces cerevisiae using tandem affinity purification and mass spectrometry. *Methods Mol Biol,* 548**,** 187-207.

BALDWIN, M. A. 2004. Protein identification by mass spectrometry: issues to be considered. *Mol Cell Proteomics,* 3**,** 1-9.

BARTOCCI, E., CORRADINI, F., MERELLI, E. & SCORTICHINI, L. 2007. BioWMS: a web-based Workflow Management System for bioinformatics. *BMC Bioinformatics,* 8 Suppl 1**,** S2.

BATEMAN, A., COIN, L., DURBIN, R., FINN, R. D., HOLLICH, V., GRIFFITHS-JONES, S., KHANNA, A., MARSHALL, M., MOXON, S., SONNHAMMER, E. L., STUDHOLME, D. J., YEATS, C. & EDDY, S. R. 2004. The Pfam protein families database. *Nucleic acids research,* 32**,** D138-41.

BELLE, A., TANAY, A., BITINCKA, L., SHAMIR, R. & O'SHEA, E. K. 2006. Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci U S A,* 103**,** 13004-9.

BETHEL, E. W., RUBEL, O., PRABHAT, WU, K. S., WEBER, G. H., PASCUCCI, V., CHILDS, H., MASCARENHAS, A., MEREDITH, J. & AHERN, S. 2009. Modern Scientific Visualization is More than Just Pretty Pictures. *Numerical Modeling of Space Plasma Flows: Astronum-2008,* 406**,** 301-316, 324.

BLAGOEV, B., KRATCHMAROVA, I., ONG, S. E., NIELSEN, M., FOSTER, L. J. & MANN, M. 2003. A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nature biotechnology,* 21**,** 315-8.

BLOW, N. 2009. Systems biology: Untangling the protein web. *Nature,* 460**,** 415-8.

BOISVERT, F. M., AHMAD, Y., GIERLINSKI, M., CHARRIERE, F., LAMOND, D., SCOTT, M., BARTON, G. & LAMOND, A. I. 2011. A quantitative spatial proteomics analysis of proteome turnover in human cells. *Molecular & cellular proteomics : MCP*.

BOISVERT, F. M., LAM, Y. W., LAMONT, D. & LAMOND, A. I. 2010. A quantitative proteomics analysis of subcellular proteome localization and changes induced by DNA damage. *Molecular & cellular proteomics : MCP,* 9**,** 457-70.

BOISVERT, F. M. & LAMOND, A. I. 2010. p53-Dependent subcellular proteome localization following DNA damage. *Proteomics,* 10**,** 4087-97.

BOISVERT, F. M., VAN KONINGSBRUGGEN, S., NAVASCUES, J. & LAMOND, A. I. 2007. The multifunctional nucleolus. *Nat Rev Mol Cell Biol,* 8**,** 574-85.

BOULON, S., AHMAD, Y., TRINKLE-MULCAHY, L., VERHEGGEN, C., COBLEY, A., GREGOR, P., BERTRAND, E., WHITEHORN, M. & LAMOND, A. I. 2010a. Establishment of a Protein Frequency Library and Its Application in the Reliable Identification of Specific Protein Interaction Partners. *Molecular & Cellular Proteomics,* 9**,** 861-879.

BOULON, S., PRADET-BALADE, B., VERHEGGEN, C., MOLLE, D., BOIREAU, S., GEORGIEVA, M., AZZAG, K., ROBERT, M. C., AHMAD, Y., NEEL, H., LAMOND, A. I. & BERTRAND, E. 2010b. HSP90 and its R2TP/Prefoldin-like cochaperone are involved in the cytoplasmic assembly of RNA polymerase II. *Mol Cell,* 39**,** 912-24.

BRADSHAW, R. A. 2005. Revised draft guidelines for proteomic data publication. *Mol Cell Proteomics,* 4**,** 1223-5.

BRADSHAW, R. A., BURLINGAME, A. L., CARR, S. & AEBERSOLD, R. 2006. Reporting protein identification data: the next generation of guidelines. *Mol Cell Proteomics,* 5**,** 787-8.

BRAND, M., RANISH, J. A., KUMMER, N. T., HAMILTON, J., IGARASHI, K., FRANCASTEL, C., CHI, T. H., CRABTREE, G. R., AEBERSOLD, R. & GROUDINE, M. 2004. Dynamic

changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. *Nat Struct Mol Biol,* 11**,** 73-80.

BROOKS, F. 1996. The computer scientist as toolsmith .2. *Communications of the Acm,* 39**,** 61-68.

BRUNNER, E., AHRENS, C. H., MOHANTY, S., BAETSCHMANN, H., LOEVENICH, S., POTTHAST, F., DEUTSCH, E. W., PANSE, C., DE LICHTENBERG, U., RINNER, O., LEE, H., PEDRIOLI, P. G., MALMSTROM, J., KOEHLER, K., SCHRIMPF, S., KRIJGSVELD, J., KREGENOW, F., HECK, A. J., HAFEN, E., SCHLAPBACH, R. & AEBERSOLD, R. 2007. A high-quality catalog of the Drosophila melanogaster proteome. *Nature biotechnology,* 25**,** 576-83.

CARD, S. K., MACKINLAY, J. D. & SHNEIDERMAN, B. 1999. *Reading In Information Visualisation: Using Vision to Think*, Academic Press.

CARR, S., AEBERSOLD, R., BALDWIN, M., BURLINGAME, A., CLAUSER, K. & NESVIZHSKII, A. 2004. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics,* 3**,** 531-3.

CARRILLO, B., YANOFSKY, C., LABOISSIERE, S., NADON, R. & KEARNEY, R. E. 2010. Methods for combining peptide intensities to estimate relative protein abundance. *Bioinformatics,* 26**,** 98-103.

CHAN, Y. L., SUZUKI, K. & WOOL, I. G. 1995. The carboxyl extensions of two rat ubiquitin fusion proteins are ribosomal proteins S27a and L40. *Biochemical and biophysical research communications,* 215**,** 682-90.

CHO, W. C. 2007. Proteomics technologies and challenges. *Genomics Proteomics Bioinformatics,* 5**,** 77-85.

CLAGUE, M. J. & URBE, S. 2010. Ubiquitin: same molecule, different degradation pathways. *Cell,* 143**,** 682-5.

CLOUTIER, P., AL-KHOURY, R., LAVALLEE-ADAM, M., FAUBERT, D., JIANG, H., POITRAS, C., BOUCHARD, A., FORGET, D., BLANCHETTE, M. & COULOMBE, B. 2009. High-resolution mapping of the protein interaction network for the human transcription machinery and affinity purification of RNA polymerase II-associated complexes. *Methods,* 48**,** 381-6.

CODD, E. F., DEAN, A. L. & ASSOCIATION FOR COMPUTING MACHINERY. SPECIAL INTEREST GROUP ON FILE DESCRIPTION AND TRANSLATION. 1971. *Data description, access and control; Proceedings of ACM-SIGFIDET Workshop, San Diego, November 11-12, 1971,* New York,, Association of Computing Machinery.

CORTHALS, G. L., WASINGER, V. C., HOCHSTRASSER, D. F. & SANCHEZ, J. C. 2000. The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis,* 21**,** 1104-15.

COTE, R. G., JONES, P., MARTENS, L., KERRIEN, S., REISINGER, F., LIN, Q., LEINONEN, R., APWEILER, R. & HERMJAKOB, H. 2007. The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics,* 8**,** 401.

COX, J. & MANN, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol,* 26**,** 1367-72.

COX, J., MATIC, I., HILGER, M., NAGARAJ, N., SELBACH, M., OLSEN, J. V. & MANN, M. 2009. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nature protocols,* 4**,** 698-705.

CRAIG, R., CORTENS, J. P. & BEAVIS, R. C. 2004. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res,* 3**,** 1234-42.

DAVIES, D. D. & HUMPHREY, T. J. 1978. Amino Acid recycling in relation to protein turnover. *Plant Physiol,* 61**,** 54-8.

DE GODOY, L. M., OLSEN, J. V., COX, J., NIELSEN, M. L., HUBNER, N. C., FROHLICH, F., WALTHER, T. C. & MANN, M. 2008. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature,* 455**,** 1251-4.

DENNIS, G., JR., SHERMAN, B. T., HOSACK, D. A., YANG, J., GAO, W., LANE, H. C. & LEMPICKI, R. A. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol,* 4**,** P3.

DESIERE, F., DEUTSCH, E. W., NESVIZHSKII, A. I., MALLICK, P., KING, N. L., ENG, J. K., ADEREM, A., BOYLE, R., BRUNNER, E., DONOHOE, S., FAUSTO, N., HAFEN, E., HOOD, L., KATZE, M. G., KENNEDY, K. A., KREGENOW, F., LEE, H., LIN, B., MARTIN, D., RANISH, J. A., RAWLINGS, D. J., SAMELSON, L. E., SHIIO, Y., WATTS, J. D., WOLLSCHEID, B., WRIGHT, M. E., YAN, W., YANG, L., YI, E. C., ZHANG, H. & AEBERSOLD, R. 2005. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol,* 6**,** R9.

DICE, J. F. & GOLDBERG, A. L. 1975. Relationship between in vivo degradative rates and isoelectric points of proteins. *Proc Natl Acad Sci U S A,* 72**,** 3893-7.

DOHERTY, M. K., HAMMOND, D. E., CLAGUE, M. J., GASKELL, S. J. & BEYNON, R. J. 2009. Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. *J Proteome Res,* 8**,** 104-12.

DOHERTY, M. K., WHITEHEAD, C., MCCORMACK, H., GASKELL, S. J. & BEYNON, R. J. 2005. Proteome dynamics in complex organisms: using stable isotopes to monitor individual protein turnover rates. *Proteomics,* 5**,** 522-33.

DOMON, B. & AEBERSOLD, R. 2006. Challenges and opportunities in proteomics data analysis. *Mol Cell Proteomics,* 5**,** 1921-6.

EDEN, E., GEVA-ZATORSKY, N., ISSAEVA, I., COHEN, A., DEKEL, E., DANON, T., COHEN, L., MAYO, A. & ALON, U. 2011. Proteome half-life dynamics in living human cells. *Science,* 331**,** 764-8.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. & BOTSTEIN, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America,* 95**,** 14863-8.

ENG, J. K., MCCORMACK, A. L. & YATES, J. R. 1994. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry,* 5**,** 976-989.

EWING, R. M., CHU, P., ELISMA, F., LI, H., TAYLOR, P., CLIMIE, S., MCBROOM-CERAJEWSKI, L., ROBINSON, M. D., O'CONNOR, L., LI, M., TAYLOR, R., DHARSEE, M., HO, Y., HEILBUT, A., MOORE, L., ZHANG, S., ORNATSKY, O., BUKHMAN, Y. V., ETHIER, M., SHENG, Y., VASILESCU, J., ABU-FARHA, M., LAMBERT, J. P., DUEWEL, H. S., STEWART, II, KUEHL, B., HOGUE, K., COLWILL, K., GLADWISH, K., MUSKAT, B., KINACH, R., ADAMS, S. L., MORAN, M. F., MORIN, G. B., TOPALOGLOU, T. & FIGEYS, D. 2007. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol,* 3**,** 89.

FAGERBERG, L., STADLER, C., SKOGS, M., HJELMARE, M., JONASSON, K., WIKING, M., ABERGH, A., UHLEN, M. & LUNDBERG, E. 2011. Mapping the Subcellular Protein Distribution in Three Human Cell Lines. *Journal of proteome research,* 10**,** 3766-3777.

FENN, J. B. 2002. Electrospray ionization mass spectrometry: How it all began. *Journal of biomolecular techniques : JBT,* 13**,** 101-18.

FENN, J. B., MANN, M., MENG, C. K., WONG, S. F. & WHITEHOUSE, C. M. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science,* 246**,** 64-71.

FENYO, D. & BEAVIS, R. C. 2003. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem,* 75**,** 768-74.

FEW, S. 2007. Data Visualization - Past, Present, And Future. Available: http://www.perceptualedge.com/articles/Whitepapers/Data_Visualization.pdf [Accessed 8th December 2010].

FOSTER, L. J., RUDICH, A., TALIOR, I., PATEL, N., HUANG, X., FURTADO, L. M., BILAN, P. J., MANN, M. & KLIP, A. 2006. Insulin-dependent interactions of proteins with GLUT4 revealed through stable isotope labeling by amino acids in cell culture (SILAC). *Journal of proteome research,* 5**,** 64-75.

GARLICK, P. J. & MILLWARD, D. J. 1972. An appraisal of techniques for the determination of protein turnover in vivo. *Proc Nutr Soc,* 31**,** 249-55.

GERSHON, D. 2003. Proteomics technologies: probing the proteome. *Nature,* 424**,** 581-7.

GREENBAUM, D., COLANGELO, C., WILLIAMS, K. & GERSTEIN, M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol,* 4**,** 117.

GYGI, S. P., RIST, B., GERBER, S. A., TURECEK, F., GELB, M. H. & AEBERSOLD, R. 1999a. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature biotechnology,* 17**,** 994-9.

GYGI, S. P., ROCHON, Y., FRANZA, B. R. & AEBERSOLD, R. 1999b. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol,* 19**,** 1720-30.

HANASH, S. & CELIS, J. E. 2002. The Human Proteome Organization: a mission to advance proteome knowledge. *Mol Cell Proteomics,* 1**,** 413-4.

HINKSON, I. V. & ELIAS, J. E. 2011. The dynamic state of protein turnover: It's about time. *Trends in cell biology,* 21**,** 293-303.

HOON, S., RATNAPU, K. K., CHIA, J. M., KUMARASAMY, B., JUGUANG, X., CLAMP, M., STABENAU, A., POTTER, S., CLARKE, L. & STUPKA, E. 2003. Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res,* 13**,** 1904-15.

HU, R. G., BROWER, C. S., WANG, H., DAVYDOV, I. V., SHENG, J., ZHOU, J., KWON, Y. T. & VARSHAVSKY, A. 2006. Arginyltransferase, its specificity, putative substrates, bidirectional promoter, and splicing-derived isoforms. *J Biol Chem,* 281**,** 32559-73.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* 4**,** 44-57.

JOHNSON, C. 2004. Top scientific visualization research problems. *Ieee Computer Graphics and Applications,* 24**,** 13-17.

JONES, P., COTE, R. G., MARTENS, L., QUINN, A. F., TAYLOR, C. F., DERACHE, W., HERMJAKOB, H. & APWEILER, R. 2006. PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res,* 34**,** D659-63.

JUNGBLUT, P. R., HOLZHUTTER, H. G., APWEILER, R. & SCHLUTER, H. 2008. The speciation of the proteome. *Chemistry Central journal,* 2**,** 16.

KARAS, M. & HILLENKAMP, F. 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem,* 60**,** 2299-301.

KEARNEY, P. & THIBAULT, P. 2003. Bioinformatics meets proteomics--bridging the gap between mass spectrometry data analysis and cell biology. *J Bioinform Comput Biol,* 1**,** 183-200.

KELLER, A., ENG, J., ZHANG, N., LI, X. J. & AEBERSOLD, R. 2005. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol,* 1**,** 2005 0017.

KOHN, D., MURRELL, G., PARKER, J. & WHITEHORN, M. 2005. What Henslow taught Darwin. *Nature,* 436**,** 643-645.

KUMAR, C. & MANN, M. 2009. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett,* 583**,** 1703-12.

LAM, Y. W., LAMOND, A. I., MANN, M. & ANDERSEN, J. S. 2007. Analysis of nucleolar protein dynamics reveals the nuclear degradation of ribosomal proteins. *Curr Biol,* 17**,** 749-60.

LAMBERT, G. N. 1984. A Comparative-Study of System Response-Time on Program Developer Productivity. *Ibm Systems Journal,* 23**,** 36-43.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, N., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LANE, D. & LEVINE, A. 2010. p53 Research: the past thirty years and the next thirty years. *Cold Spring Harb Perspect Biol,* 2**,** a000893.

LARANCE, M., BAILLY, A. P., POURKARIMI, E., HAY, R. T., BUCHANAN, G., COULTHURST, S., XIRODIMAS, D. P., GARTNER, A. & LAMOND, A. I. 2011. Stable-isotope labeling with amino acids in nematodes. *Nature methods,* 8**,** 849-51.

LETUNIC, I., COPLEY, R. R., SCHMIDT, S., CICCARELLI, F. D., DOERKS, T., SCHULTZ, J., PONTING, C. P. & BORK, P. 2004. SMART 4.0: towards genomic data integration. *Nucleic acids research,* 32**,** D142-4.

LEUNG, A. K., TRINKLE-MULCAHY, L., LAM, Y. W., ANDERSEN, J. S., MANN, M. & LAMOND, A. I. 2006. NOPdb: Nucleolar Proteome Database. *Nucleic Acids Res,* 34**,** D218-20.

LUHN, H. P. 1958. A business intelligence system. *IBM J. Res. Dev.,* 2**,** 314-319.

LUNDBERG, E., FAGERBERG, L., KLEVEBRING, D., MATIC, I., GEIGER, T., COX, J., ALGENAS, C., LUNDEBERG, J., MANN, M. & UHLEN, M. 2010. Defining the transcriptome and proteome in three functionally different human cell lines. *Mol Syst Biol,* 6**,** 450.

MACRAE, J. I. & FERGUSON, M. A. 2005. A robust and selective method for the quantification of glycosylphosphatidylinositols in biological samples. *Glycobiology,* 15**,** 131-8.

MAIOLICA, A., CITTARO, D., BORSOTTI, D., SENNELS, L., CIFERRI, C., TARRICONE, C., MUSACCHIO, A. & RAPPSILBER, J. 2007. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Molecular & cellular proteomics : MCP,* 6**,** 2200-11.

MANN, M. 2006. Functional and quantitative proteomics using SILAC. *Nature reviews. Molecular cell biology,* 7**,** 952-8.

MARTENS, L. 2006. *Novel bioinformatics tools assisting tergeted peptide-centric proteomics and global proteomics data dissemination.* Doctorate (PhD) in Sciences: Biotechnology, Ghent University.

MARTIN, D. M., NETT, I. R., VANDERMOERE, F., BARBER, J. D., MORRICE, N. A. & FERGUSON, M. A. 2010. Prophossi: automating expert validation of

phosphopeptide-spectrum matches from tandem mass spectrometry. *Bioinformatics,* 26**,** 2153-9.

MATIC, I., VAN HAGEN, M., SCHIMMEL, J., MACEK, B., OGG, S. C., TATHAM, M. H., HAY, R. T., LAMOND, A. I., MANN, M. & VERTEGAAL, A. C. 2008. In vivo identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an in vitro to in vivo strategy. *Molecular & cellular proteomics : MCP,* 7**,** 132-44.

MATLIN, A. J., CLARK, F. & SMITH, C. W. 2005. Understanding alternative splicing: towards a cellular code. *Nature reviews. Molecular cell biology,* 6**,** 386-98.

MCCORMICK, B. H., DEFANTI, T. A. & BROWN, M. D. 1987. Visualization in Scientific Computing. *Computer Graphics,* 27.

MCDONALD, W. H., TABB, D. L., SADYGOV, R. G., MACCOSS, M. J., VENABLE, J., GRAUMANN, J., JOHNSON, J. R., COCIORVA, D. & YATES, J. R., 3RD 2004. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun Mass Spectrom,* 18**,** 2162-8.

MEAD, J. A., SHADFORTH, I. P. & BESSANT, C. 2007. Public proteomic MS repositories and pipelines: available tools and biological applications. *Proteomics,* 7**,** 2769-86.

MILNER, E., BARNEA, E., BEER, I. & ADMON, A. 2006. The turnover kinetics of major histocompatibility complex peptides of human cancer cells. *Mol Cell Proteomics,* 5**,** 357-65.

MULDER, N. J., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BARRELL, D., BATEMAN, A., BINNS, D., BISWAS, M., BRADLEY, P., BORK, P., BUCHER, P., COPLEY, R. R., COURCELLE, E., DAS, U., DURBIN, R., FALQUET, L., FLEISCHMANN, W., GRIFFITHS-JONES, S., HAFT, D., HARTE, N., HULO, N., KAHN, D., KANAPIN, A., KRESTYANINOVA, M., LOPEZ, R., LETUNIC, I., LONSDALE, D., SILVENTOINEN, V., ORCHARD, S. E., PAGNI, M., PEYRUC, D., PONTING, C. P., SELENGUT, J. D., SERVANT, F., SIGRIST, C. J.,

VAUGHAN, R. & ZDOBNOV, E. M. 2003. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic acids research,* 31**,** 315-8.

MUNZNER, T. 2000. *Interactive Visualization of Large Graphs and Networks.* Stanford University.

NEERINCX, P. B. & LEUNISSEN, J. A. 2005. Evolution of web services in bioinformatics. *Brief Bioinform,* 6**,** 178-88.

NESVIZHSKII, A. I. & AEBERSOLD, R. 2004. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov Today,* 9**,** 173-81.

NESVIZHSKII, A. I., VITEK, O. & AEBERSOLD, R. 2007. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods,* 4**,** 787-97.

NETT, I. R., DAVIDSON, L., LAMONT, D. & FERGUSON, M. A. 2009a. Identification and specific localization of tyrosine-phosphorylated proteins in Trypanosoma brucei. *Eukaryotic cell,* 8**,** 617-26.

NETT, I. R., MARTIN, D. M., MIRANDA-SAAVEDRA, D., LAMONT, D., BARBER, J. D., MEHLERT, A. & FERGUSON, M. A. 2009b. The phosphoproteome of bloodstream form Trypanosoma brucei, causative agent of African sleeping sickness. *Molecular & cellular proteomics : MCP,* 8**,** 1527-38.

NETT, I. R., MEHLERT, A., LAMONT, D. & FERGUSON, M. A. 2010. Application of electrospray mass spectrometry to the structural determination of glycosylphosphatidylinositol membrane anchors. *Glycobiology,* 20**,** 576-85.

NEUBAUER, G., KING, A., RAPPSILBER, J., CALVIO, C., WATSON, M., AJUH, P., SLEEMAN, J., LAMOND, A. & MANN, M. 1998. Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature genetics,* 20**,** 46-50.

NILSSON, T., MANN, M., AEBERSOLD, R., YATES, J. R., 3RD, BAIROCH, A. & BERGERON, J. J. 2010. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nature methods,* 7**,** 681-5.

O'DONOVAN, C., MARTIN, M. J., GATTIKER, A., GASTEIGER, E., BAIROCH, A. & APWEILER, R. 2002. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform,* 3**,** 275-84.

OHSUMI, Y. 2006. Protein turnover. *IUBMB Life,* 58**,** 363-9.

OHTA, S., BUKOWSKI-WILLS, J. C., SANCHEZ-PULIDO, L., ALVES FDE, L., WOOD, L., CHEN, Z. A., PLATANI, M., FISCHER, L., HUDSON, D. F., PONTING, C. P., FUKAGAWA, T., EARNSHAW, W. C. & RAPPSILBER, J. 2010. The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell,* 142**,** 810-21.

OINN, T., ADDIS, M., FERRIS, J., MARVIN, D., SENGER, M., GREENWOOD, M., CARVER, T., GLOVER, K., POCOCK, M. R., WIPAT, A. & LI, P. 2004. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics,* 20**,** 3045-54.

ONG, S. E., BLAGOEV, B., KRATCHMAROVA, I., KRISTENSEN, D. B., STEEN, H., PANDEY, A. & MANN, M. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics,* 1**,** 376-86.

ONG, S. E., KRATCHMAROVA, I. & MANN, M. 2003. Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *Journal of proteome research,* 2**,** 173-81.

ONG, S. E. & MANN, M. 2005. Mass spectrometry-based proteomics turns quantitative. *Nature chemical biology,* 1**,** 252-62.

ONG, S. E. & MANN, M. 2006. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat Protoc,* 1**,** 2650-60.

ORCHARD, S., HERMJAKOB, H., TAYLOR, C. F., POTTHAST, F., JONES, P., ZHU, W., JULIAN, R. K., JR. & APWEILER, R. 2005. Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17-20th April 2005). *Proteomics,* 5**,** 3552-5.

PAUL, W. 1990. Electromagnetic traps for charged and neutral particles. *Rev. Mod. Phys.,* 62**,** 531-540.

PEDRIOLI, P. G., ENG, J. K., HUBLEY, R., VOGELZANG, M., DEUTSCH, E. W., RAUGHT, B., PRATT, B., NILSSON, E., ANGELETTI, R. H., APWEILER, R., CHEUNG, K., COSTELLO, C. E., HERMJAKOB, H., HUANG, S., JULIAN, R. K., KAPP, E., MCCOMB, M. E., OLIVER, S. G., OMENN, G., PATON, N. W., SIMPSON, R., SMITH, R., TAYLOR, C. F., ZHU, W. & AEBERSOLD, R. 2004. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol,* 22**,** 1459-66.

PERKINS, D. N., PAPPIN, D. J., CREASY, D. M. & COTTRELL, J. S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis,* 20**,** 3551-67.

PIGGEE, C. 2008. LIMS and the art of MS proteomics. *Anal Chem,* 80**,** 4801-6.

POST, F. H., NIELSON, G. M. & BONNEAU, G.-P. 2003. *Data Visualization: The State of the Art*, Kluwer Academic Publishers.

PRATT, J. M., PETTY, J., RIBA-GARCIA, I., ROBERTSON, D. H., GASKELL, S. J., OLIVER, S. G. & BEYNON, R. J. 2002. Dynamics of protein turnover, a missing dimension in proteomics. *Mol Cell Proteomics,* 1**,** 579-91.

PRINCE, J. T., CARLSON, M. W., WANG, R., LU, P. & MARCOTTE, E. M. 2004. The need for a public proteomics repository. *Nat Biotechnol,* 22**,** 471-2.

RANISH, J. A., BRAND, M. & AEBERSOLD, R. 2007. Using stable isotope tagging and mass spectrometry to characterize protein complexes and to detect changes in their composition. *Methods Mol Biol,* 359**,** 17-35.

RAPPSILBER, J. 2011. The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *Journal of structural biology,* 173**,** 530-40.

RAPPSILBER, J. & MANN, M. 2002a. Is mass spectrometry ready for proteome-wide protein expression analysis? *Genome biology,* 3**,** COMMENT2008.

RAPPSILBER, J. & MANN, M. 2002b. What does it mean to identify a protein in proteomics? *Trends in biochemical sciences,* 27**,** 74-8.

RAPPSILBER, J., SINIOSSOGLOU, S., HURT, E. C. & MANN, M. 2000. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Analytical chemistry,* 72**,** 267-75.

RECHSTEINER, M. & ROGERS, S. W. 1996. PEST sequences and regulation by proteolysis. *Trends Biochem Sci,* 21**,** 267-71.

RHYNE, T. M. 2003. Does the difference between information and scientific visualization really matter? *Ieee Computer Graphics and Applications,* 23**,** 6-8.

RIGAUT, G., SHEVCHENKO, A., RUTZ, B., WILM, M., MANN, M. & SERAPHIN, B. 1999. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol,* 17**,** 1030-2.

ROGERS, S., WELLS, R. & RECHSTEINER, M. 1986. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science,* 234**,** 364-8.

ROHLFF, C. 2004. New approaches towards integrated proteomic databases and depositories. *Expert Rev Proteomics,* 1**,** 267-74.

ROSS, P. L., HUANG, Y. N., MARCHESE, J. N., WILLIAMSON, B., PARKER, K., HATTAN, S., KHAINOVSKI, N., PILLAI, S., DEY, S., DANIELS, S., PURKAYASTHA, S., JUHASZ, P., MARTIN, S., BARTLET-JONES, M., HE, F., JACOBSON, A. & PAPPIN, D. J. 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics : MCP,* 3**,** 1154-69.

SCHLEGEL, K. 2008. Emerging Technologies Will Drive Self-Service Business Intelligence. Gartner.

SCHULZE, W. X. & MANN, M. 2004. A novel proteomic screen for peptide-protein interactions. *J Biol Chem,* 279**,** 10756-64.

SCHWANHAUSSER, B., BUSSE, D., LI, N., DITTMAR, G., SCHUCHHARDT, J., WOLF, J., CHEN, W. & SELBACH, M. 2011. Global quantification of mammalian gene expression control. *Nature,* 473**,** 337-42.

SCHWANHAUSSER, B., GOSSEN, M., DITTMAR, G. & SELBACH, M. 2009. Global analysis of cellular protein translation by pulsed SILAC. *Proteomics,* 9**,** 205-9.

SELBACH, M. & MANN, M. 2006. Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nat Methods,* 3**,** 981-3.

SHAH, S. P., HE, D. Y., SAWKINS, J. N., DRUCE, J. C., QUON, G., LETT, D., ZHENG, G. X., XU, T. & OUELLETTE, B. F. 2004. Pegasys: software for executing and integrating analyses of biological sequences. *BMC Bioinformatics,* 5**,** 40.

SHNEIDERMAN, B. The eyes have it: A task by data type taxonomy for information visualizations. IEEE Symposium on Visual Languages, 1996. 336-345.

SMALLMON, T. R. & GANJEI, J. K. 2004. The benefits of a LIMS in proteomics. *LIMS/Letter.*

SQUIRES, G. 1998. Francis Aston and the mass spectrograph. *Journal of the Chemical Society, Dalton Transactions***,** 3893-3900.

SUNDQVIST, A., LIU, G., MIRSALIOTIS, A. & XIRODIMAS, D. P. 2009. Regulation of nucleolar signalling to p53 through NEDDylation of L11. *EMBO Rep,* 10**,** 1132-9.

TACKETT, A. J., DEGRASSE, J. A., SEKEDAT, M. D., OEFFINGER, M., ROUT, M. P. & CHAIT, B. T. 2005. I-DIRT, a general method for distinguishing between specific and nonspecific protein interactions. *J Proteome Res,* 4**,** 1752-6.

TANAKA, K., WAKI, H., IDO, Y., AKITA, S., YOSHIDA, Y., YOSHIDA, T. & MATSUO, T. 1988. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid communications in mass spectrometry,* 2**,** 151-153.

TANG, F., CHUA, C. L., HO, L. Y., LIM, Y. P., ISSAC, P. & KRISHNAN, A. 2005. Wildfire: distributed, Grid-enabled workflow construction and execution. *BMC Bioinformatics,* 6**,** 69.

TEN HAVE, S., BOULON, S., AHMAD, Y. & LAMOND, A. I. 2011. Mass spectrometry-based immuno-precipitation proteomics - the user's guide. *Proteomics,* 11**,** 1153-9.

THOMSON, J. J. 2010. Cathode Rays (Reprinted from Philosophical Magazine Series 5, vol 44, pg 293-316, 1897). *Philosophical Magazine,* 90**,** 25-29.

TRINKLE-MULCAHY, L., ANDERSEN, J., LAM, Y. W., MOORHEAD, G., MANN, M. & LAMOND, A. I. 2006. Repo-Man recruits PP1 gamma to chromatin and is essential for cell viability. *J Cell Biol,* 172**,** 679-92.

TRINKLE-MULCAHY, L., BOULON, S., LAM, Y. W., URCIA, R., BOISVERT, F. M., VANDERMOERE, F., MORRICE, N. A., SWIFT, S., ROTHBAUER, U., LEONHARDT, H. & LAMOND, A. 2008a. Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J Cell Biol,* 183**,** 223-39.

TRINKLE-MULCAHY, L., BOULON, S., LAM, Y. W., URCIA, R., BOISVERT, F. M., VANDERMOERE, F., MORRICE, N. A., SWIFT, S., ROTHBAUER, U., LEONHARDT, H. & LAMOND, A. 2008b. Identifying specific protein interaction partners using

quantitative mass spectrometry and bead proteomes. *The Journal of cell biology,* 183**,** 223-39.

TRINKLE-MULCAHY, L., BOULON, S. V., LAM, Y. W., URCIA, R., BOISVERT, F. O.-M., VANDERMOERE, F., MORRICE, N. A., SWIFT, S., ROTHBAUER, U., LEONHARDT, H. & LAMOND, A. 2008c. Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *The Journal of Cell Biology,* 183**,** 223-239.

TUFTE, E. R. 1983. *The Visual Display of Quantitative Information,* Cheshire, Graphics Press.

TUKEY, J. W. 1977. *Exploratory data analysis,* Reading, Mass., Addison-Wesley Pub. Co.

TYERS, M. & MANN, M. 2003. From genomics to proteomics. *Nature,* 422**,** 193-7.

VARSHAVSKY, A. 1992. The N-End Rule. *Cell,* 69**,** 725-735.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J., GABOR MIKLOS, G. L., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, J., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z., DI FRANCESCO, V., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W., GONG, F., GU, Z., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R. R., KE, Z., KETCHUM, K. A., LAI, Z., LEI, Y., LI, Z., LI, J., LIANG, Y., LIN, X., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B., SUN, J., WANG, Z., WANG, A., WANG, X., WANG, J., WEI, M.,

WIDES, R., XIAO, C., YAN, C., et al. 2001. The sequence of the human genome. *Science,* 291**,** 1304-51.

VERMEULEN, M., HUBNER, N. C. & MANN, M. 2008. High confidence determination of specific protein-protein interactions using quantitative mass spectrometry. *Curr Opin Biotechnol,* 19**,** 331-7.

VERTEGAAL, A. C., ANDERSEN, J. S., OGG, S. C., HAY, R. T., MANN, M. & LAMOND, A. I. 2006. Distinct and overlapping sets of SUMO-1 and SUMO-2 target proteins revealed by quantitative proteomics. *Molecular & cellular proteomics : MCP,* 5**,** 2298-310.

WALTHER, T. C. & MANN, M. 2010a. Mass spectrometry-based proteomics in cell biology. *The Journal of cell biology,* 190**,** 491-500.

WALTHER, T. C. & MANN, M. 2010b. Mass spectrometry-based proteomics in cell biology. *J Cell Biol,* 190**,** 491-500.

WESTMAN, B. J., VERHEGGEN, C., HUTTEN, S., LAM, Y. W., BERTRAND, E. & LAMOND, A. I. 2010. A proteomic screen for nucleolar SUMO targets shows SUMOylation modulates the function of Nop5/Nop58. *Molecular cell,* 39**,** 618-31.

WILKINS, M. R., APPEL, R. D., VAN EYK, J. E., CHUNG, M. C., GORG, A., HECKER, M., HUBER, L. A., LANGEN, H., LINK, A. J., PAIK, Y. K., PATTERSON, S. D., PENNINGTON, S. R., RABILLOUD, T., SIMPSON, R. J., WEISS, W. & DUNN, M. J. 2006. Guidelines for the next 10 years of proteomics. *Proteomics,* 6**,** 4-8.

WILKINSON, M. D., GESSLER, D., FARMER, A. & STEIN, L. The BioMOBY project explores open-source, simple, extensible protocols for enabling biological database interoperability. Proceedings of the Virtual Conference on Genomics and Bioinformatics, 2003. 17-27.

WILSON, I. B., O'DONNELL, N., ALLEN, S., MEHLERT, A. & FERGUSON, M. A. 1999. Typing of Leishmania lipophosphoglycans by electrospray mass spectrometry. *Molecular and biochemical parasitology,* 100**,** 207-15.

YAN, Y., PHAN, L., YANG, F., TALPAZ, M., YANG, Y., XIONG, Z., NG, B., TIMCHENKO, N. A., WU, C. J., RITZ, J., WANG, H. & YANG, X. F. 2004. A novel mechanism of alternative promoter and splicing regulates the epitope generation of tumor antigen CML66-L. *Journal of immunology,* 172**,** 651-60.

YATES, J. R., 3RD, GILCHRIST, A., HOWELL, K. E. & BERGERON, J. J. 2005. Proteomics of organelles and large cellular structures. *Nature reviews. Molecular cell biology,* 6**,** 702-14.

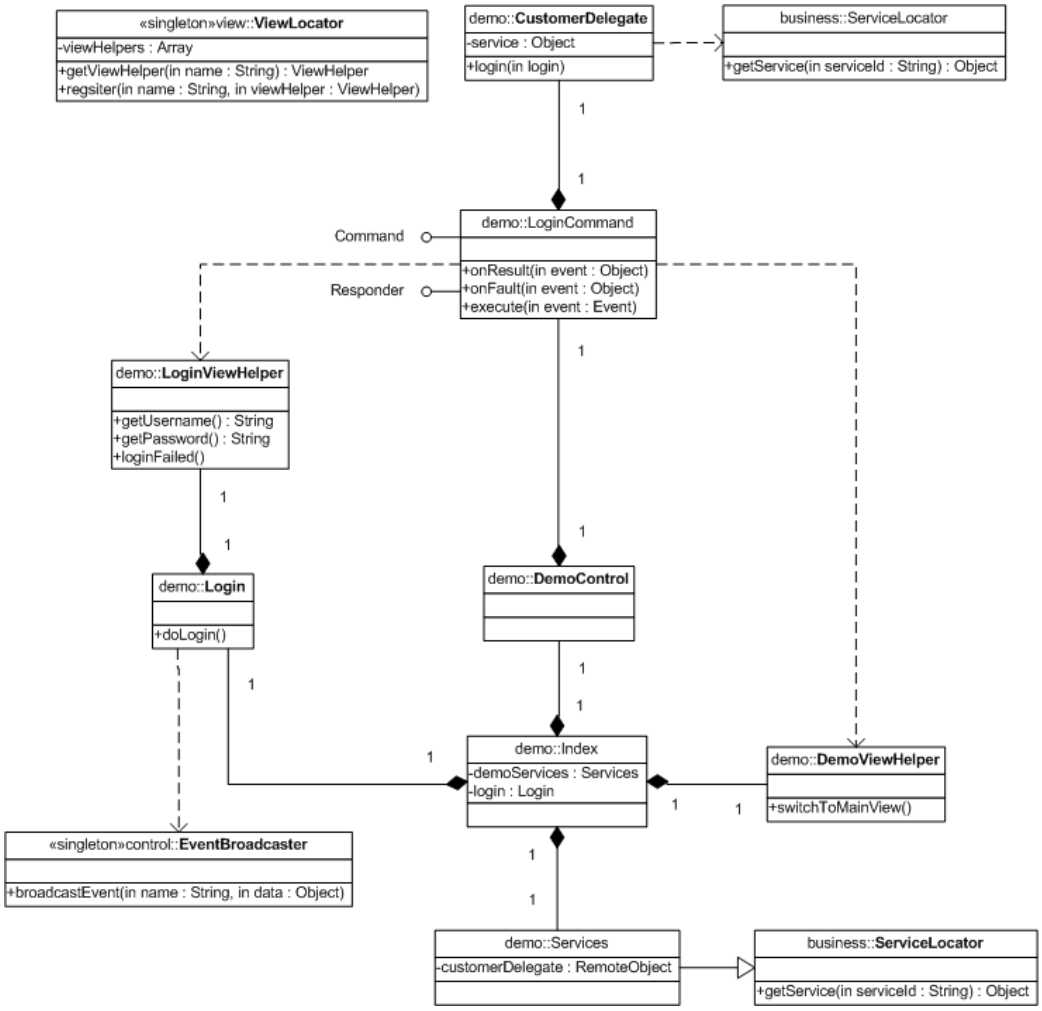
YEH, R. K. 2006. Visualization Techniques for Data Mining in Business Context - A Comparative Analysis. *Decision Sciences Institute, Southwest Region.* Oklahoma City.

YEN, H. C., XU, Q., CHOU, D. M., ZHAO, Z. & ELLEDGE, S. J. 2008. Global protein stability profiling in mammalian cells. *Science,* 322**,** 918-23.

Appendices

A. Cairngorm Framework

The following class diagram depicts the Cairngorm Framework. This framework describes how to design and implement solutions. Hence it allows developers to easily analyse and understand code written by other developers. Furthermore, this framework ensures the design of the Vision application remains maintainable and extendable.



B. Metadata Definition

The following metadata is collected with regards to mass spectrometry submissions.

General Details

Field	Definition
Keywords	Specify up to three words associated with the experiment.
Cost Centre	Cost centre number to be used by MS facility for billing.
Organism	Organism of the cell type used in the experiment.
Cell Type	Main cell type used in the experiment.
Treatment Type	Specify whether drug or stress treatment(s) were used.
Treatment	Details of treatment(s) applied to cells.
Enzyme	Enzyme used in the experiment.
Digestion	Type of digestion used in the experiment.
Additional Information	Any additional details that a user may want to specify.

Mass Spectrometry Details

Field	Definition
Instrument	The instrument you would like your samples to be run on.
Reagent	Reagent used during alkylation to reduce proteins, e.g. IAA. Note this will change the mass of your peptides.
Sample Composition	Solution used to purify peptides, e.g. 1% FA. This should be an analytical grade acid solution, such as 1% Formic Acid or 0.1% Trifluoroacetic Acid. Please consult the MS facility if you intend to use a different solution.
Peptide Cleanup	Method used for peptide cleanup. Peptide cleanup should be carried out on ALL samples.
Sample Volume	Total volume of sample being submitted to the MS facility in microlitres.
Injection Volume	Volume of sample to inject on the mass spectrometer in microlitres.
Analysis Type	Type of MS analysis.
Quant Type	Type of quantification used, if any.
Run Length	Length of run for each MS sample. Suggested times include: 40mins for purified, simple, single or small complex protein sample ID; 100 mins for complex fractionated samples; and 180mins for complex protein samples.
96 Well Plate	Indicate whether a 96 well plate will be used for the sample submission.

SILAC Details

Field	Definition
Medium Labels	The labelling used in the medium SILAC label.
Heavy Labels	The labelling used in the heavy SILAC label.

Pull-Down Details

Field	Definition
Bait	Type of target protein to be pulled down.
Tag Type	Labelling of protein, if any.
Bead Type	Solid matrix used.
Bead	Bead used to pull-down protein, if any.
Antibody Type	Antibody used in experiment.
Buffers	Buffers used during pull down of protein. Note: IP buffer specifics affect which proteins are pulled down in each experiment, i.e. high salt percentage will mean less proteins are pulled down.
Pre-Clearing Time	Pre-clearing is a step that can be included where the protein solution is incubated for a short time with beads to reduce non-specific protein binding.
Incubation Time	Incubation time with the antibody.

Gel Lane/Solution Sample Details

Field	Definition
Cell Extract	Specify the cell extract of the lane/samples.
Derived Cell Type	Relevant mutant or genetically modified cell types used to express a specific protein.
Cell Cycle Stage	Specify the cell cycle phase of cells in the lane/samples.
Volume	Micrograms per microlitre of protein loaded onto gel/in-solution.
Replicate	Replicate type of sample, if applicable.
Samples	Number of fractions the gel lane was separated into.
IP Protein	Protein(s) pulled down in lane/sample.
Light Label	Description of the lane/samples in light media.
Medium Label	Description of the lane/samples in medium media
Heavy Label	Description of the lane/samples in heavy media
Column One Type	Chromatography column one type.
Column One Buffer	Chromatography column one buffer description.
Column Two Type	Chromatography column two type.
Column Two Buffer	Chromatography column two buffer description.
Column One Fractions	Number of fractions for chromatography column one.
Column Two Fractions	Number of fractions for chromatography column two.
Fraction Volume	Volume of fractions on microliters.
Volume Injected	Volume injected in microliters.

C. N-End Rule Evaluation

This table refers to data generated during the study described in Chapter 6: Spatial Localisation & Turnover Analyses

Chapter 6: Spatial Localisation & Turnover Analyses. Analysis was carried out to determine whether the amino acids at the N-terminus or C- terminus of a protein sequence could be related to the turnover or half-life of the protein, as suggested by the N-end rule.

This analysis considered all turnover proteins that start with a methionine at the N-terminus. The first table (see Table 7) shows the average half-life measured for each amino acid at the first ten N-terminal positions and Table 8, similarly shows the average half-life measured for each amino acid at the last ten C-terminal positions. From these tables we can conclude that half-life is not affected by the amino acid identity at the N-terminus or C- terminus. Furthermore, Table 9 and Table 10 show that the average turnover rate is also not affected by the amino acid identity at either the N-terminus or C-terminus.

N - End	AVG(Half Life)									
Amino Acid	1st Position	2nd Position	3rd Position	4th Position	5th Position	6th Position	7th Position	8th Position	9th Position	10th Position
A		22.5	22.4	22.1	22.3	22.7	22.3	21.7	21.7	22.0
R		20.4	23.0	21.8	21.0	22.1	21.1	22.3	21.7	22.0
N		20.9	23.0	20.1	22.2	20.6	22.1	21.5	21.3	21.4
D		20.3	21.9	22.5	20.9	22.0	22.2	20.7	21.9	21.9
C		25.2	20.8	22.8	22.9	21.9	21.2	20.5	21.5	22.4
Q		20.9	21.7	21.8	21.0	22.6	21.9	22.1	22.7	23.8
E		19.9	22.7	22.0	21.5	20.9	21.0	20.7	21.8	22.7
G		21.1	21.4	22.4	21.0	21.2	21.6	22.0	21.7	21.9
H		21.7	21.7	22.0	23.8	21.9	20.5	19.7	20.4	19.2
I		22.4	21.4	22.0	21.8	21.6	23.7	22.9	21.9	21.5
L		24.6	21.3	22.7	23.0	22.3	21.7	23.0	22.9	21.9
K		21.8	21.7	21.5	21.9	20.4	20.7	21.6	20.8	21.4
M	21.8	20.7	22.6	19.9	20.2	20.8	19.2	21.0	21.2	21.6
F		22.5	25.5	20.0	21.0	22.4	22.1	22.0	21.1	19.7
P		22.6	20.8	20.9	20.9	21.5	22.3	21.6	21.3	20.9
S		21.9	20.8	21.9	21.6	21.8	21.9	21.6	21.8	21.7
T		20.1	20.4	21.4	21.5	21.0	22.0	21.2	21.3	21.8
W		19.2	24.0	21.1	21.0	20.8	21.3	21.3	21.0	22.7
Y		24.4	21.2	21.5	23.6	23.5	25.8	22.1	21.7	22.5
V		21.9	20.6	22.2	23.6	22.2	21.7	22.1	22.2	21.3

Table 7: Average half-life for each amino acid at the first ten N-terminal positions.

C - End	AVG(Half Life)									
Amino Acid	1st Position	2nd Position	3rd Position	4th Position	5th Position	6th Position	7th Position	8th Position	9th Position	10th Position
A	22.7	20.7	22.8	21.4	23.1	22.1	21.4	22.1	21.3	21.6
R	21.6	22.5	21.4	21.4	21.4	22.2	22.2	21.7	22.2	21.7
N	22.4	21.3	21.1	21.6	21.7	22.4	22.7	20.1	23.3	22.4
D	21.8	20.0	21.0	21.3	20.3	21.9	20.6	21.4	23.5	21.0
C	21.8	22.1	21.3	18.8	22.2	20.1	20.3	21.0	21.0	18.3
Q	21.3	22.4	22.2	23.6	22.4	21.8	20.7	21.4	21.6	22.2
E	21.6	22.4	21.3	21.6	23.5	21.2	22.6	21.8	22.2	21.6
G	21.8	21.7	22.7	21.1	22.4	22.7	21.7	23.2	22.7	21.7
H	21.9	21.2	23.1	24.3	22.1	21.0	22.8	20.6	21.0	20.9
I	21.1	21.0	20.7	22.4	23.2	21.7	21.5	21.7	23.4	22.1
L	21.2	21.0	21.8	21.7	22.1	21.8	22.4	22.4	22.4	21.2
K	22.6	22.7	22.5	22.2	21.7	21.9	21.9	21.4	21.2	22.6
M	20.5	24.5	19.6	19.1	20.9	24.0	22.2	21.5	22.6	22.0
F	22.5	23.6	22.6	21.3	21.9	21.6	22.8	21.7	21.7	21.0
P	21.9	22.0	22.3	21.8	20.4	22.0	21.9	22.1	21.8	23.0
S	22.0	21.3	22.5	21.6	21.1	21.3	20.9	21.4	20.2	20.9
T	21.1	22.2	21.1	22.7	21.1	20.7	21.5	22.6	22.3	22.5
W	20.3	22.6	20.2	22.1	21.5	22.8	21.2	20.9	21.9	19.4
Y	21.4	21.9	21.7	21.5	21.6	22.1	23.5	24.3	19.7	21.3
V	22.8	22.5	21.5	24.0	20.8	22.1	21.9	21.8	20.7	24.2

Table 8: Average half-life for each amino acid at the last ten C-terminal positions.

N - End	AVG(Turnover)									
Amino Acid	1st Position	2nd Position	3rd Position	4th Position	5th Position	6th Position	7th Position	8th Position	9th Position	10th Position
A		19.7	19.7	19.2	20.0	19.9	19.8	19.6	19.1	19.0
R		18.5	19.6	19.0	18.6	19.2	18.9	19.4	19.2	19.3
N		19.0	18.6	18.4	19.3	19.8	19.4	19.7	19.6	19.1
D		18.3	18.9	19.3	19.3	19.6	19.0	19.0	19.8	19.6
C		19.2	18.6	21.3	19.8	18.7	19.3	19.4	18.9	19.9
Q		19.0	19.0	19.2	19.6	19.5	18.4	19.1	19.0	19.6
E		18.3	20.3	19.7	19.2	19.0	19.6	18.5	19.7	19.1
G		19.2	19.7	19.6	18.8	19.1	19.3	19.5	19.5	19.8
H		18.9	20.0	20.8	20.4	20.2	20.3	19.0	18.8	18.9
I		19.8	19.6	19.9	19.1	19.3	20.1	19.7	20.0	19.2
L		20.5	19.5	20.1	20.0	19.6	19.1	19.7	19.5	19.4
K		19.2	19.4	19.7	19.7	18.6	19.2	19.5	18.8	19.4
M	19.3	19.5	20.1	18.2	17.1	17.8	19.7	17.9	20.6	17.9
F		20.5	20.4	18.4	19.2	20.3	18.7	19.4	19.4	18.3
P		20.2	18.5	19.0	19.1	19.0	19.8	19.5	18.7	19.0
S		19.1	18.6	19.4	19.1	18.9	19.5	19.0	19.2	19.9
T		18.5	19.0	19.2	19.0	19.4	19.3	18.6	19.5	19.9
W		19.4	19.5	18.3	18.6	17.5	17.6	19.6	18.4	20.6
Y		22.0	20.3	19.1	20.7	18.0	20.1	19.7	20.5	20.1
V		19.0	18.6	19.1	19.8	20.6	19.4	19.7	19.7	19.0

Table 9: Average turnover for each amino acid at the first ten N-terminal positions.

C - End	AVG(Turnover)									
Amino Acid	1st Position	2nd Position	3rd Position	4th Position	5th Position	6th Position	7th Position	8th Position	9th Position	10th Position
A	19.9	19.2	19.8	19.2	19.9	20.2	19.2	19.9	19.3	19.3
R	19.6	20.1	19.3	19.1	19.2	18.9	19.5	19.5	19.1	19.5
N	19.7	19.2	18.6	19.2	19.1	19.6	19.6	19.1	18.7	19.5
D	19.5	18.7	19.9	19.0	18.5	18.6	19.1	19.6	19.9	18.9
C	19.3	18.8	18.7	17.9	20.0	19.0	19.8	19.4	20.1	17.9
Q	19.0	19.9	19.2	20.0	20.0	19.9	19.5	19.3	19.6	19.3
E	19.4	19.8	19.4	19.9	19.7	19.5	19.9	19.5	20.4	19.5
G	19.4	18.9	19.3	19.1	20.0	20.0	19.0	19.4	19.3	19.3
H	18.8	18.4	20.0	20.0	19.0	18.9	20.2	19.1	19.5	18.1
I	18.9	19.4	19.0	19.5	20.3	19.3	18.8	18.7	19.7	19.1
L	19.3	19.1	19.9	19.5	19.9	19.4	19.7	19.4	19.6	19.0
K	19.7	19.3	20.1	19.6	19.3	19.4	19.9	19.2	19.1	20.1
M	19.1	21.3	17.9	18.2	19.2	20.5	20.3	20.5	20.5	19.6
F	19.4	20.5	18.8	18.8	19.1	18.9	18.9	19.0	19.3	19.5
P	19.1	19.4	19.3	19.3	18.5	19.3	19.1	19.2	19.2	19.3
S	19.2	19.1	19.4	19.2	18.3	18.9	18.7	19.1	18.6	19.0
T	19.3	19.5	19.1	19.1	19.4	18.7	19.2	19.5	19.7	20.0
W	17.9	18.4	17.4	20.6	19.6	20.5	18.7	19.1	19.2	17.2
Y	19.4	19.6	19.4	19.1	19.9	19.6	19.6	20.4	17.8	18.6
V	20.0	19.5	19.3	20.3	19.2	19.2	19.2	19.2	19.3	21.0

Table 10: Average turnover for each amino acid at the last ten C-terminal positions.

D. Random Protein Sampling Evaluation

This table refers to data generated during the study described in Chapter 6: Spatial Localisation & Turnover Analyses.

In order to evaluate the random sampling of proteins within the study, the amino acid occurrence at the first ten N-terminal positions were calculated for the complete human proteome (see Table 11) and compared with the matrix of frequencies calculated for proteins identified in this study (see Table 12). The Pearson correlation of this comparison is 0.99 indicating that the subset of human proteins sampled in the study is highly representative of the total human proteome.

In order to ensure, this correlation was truly reflective of all data generated during this study, the matrix of frequencies was calculated for the top 10% fastest turnover proteins (see Table 13, Pearson correlation: 0.99) and the top 10% slowest turnover proteins (see Table 14, Pearson correlation: 0.98). The high correlations show that all sets of proteins within this study are highly representative of the total human proteome.

Complete Human Proteome										
Amino Acid	1st Position	2nd Position	3rd Position	4th Position	5th Position	6th Position	7th Position	8th Position	9th Position	10th Position
A	0	14765	7608	6822	6398	6646	6405	6492	6720	6371
R	0	3807	5514	5684	5663	5910	5666	5575	5286	5318
N	0	2605	2388	2417	2403	2115	2255	2101	2039	1867
D	0	4321	3405	2568	2956	3226	2822	2649	3022	2955
C	0	1088	1584	1768	1743	1884	1828	1872	2000	2149
Q	0	2180	3207	3353	3312	3349	3393	3329	3238	3132
E	0	6932	5245	5007	4719	4635	4645	4541	4417	4423
G	0	6153	5761	5456	6729	6259	5562	5983	5910	5863
H	0	1148	1689	1676	1600	1572	1688	1645	1613	1579
I	0	1681	2131	2552	2477	2472	2426	2748	2667	2772
L	0	4961	7331	8513	8006	8549	9501	9597	9946	10191
K	0	3804	3917	4211	4449	4004	4079	3763	3598	3491
M	78546	1426	1451	1502	1470	1392	1331	1574	1238	1310
F	0	1823	2434	2727	2609	2618	2687	2695	3014	3108
P	0	4308	5616	5989	5202	5438	5787	5722	5307	5710
S	0	8157	8033	7293	7107	7039	6954	6361	6656	6746
T	0	3604	4658	4088	3996	4033	4189	4121	4023	3955
W	0	1226	1347	1334	1462	1341	1335	1383	1318	1370
Y	0	1015	1151	1413	1383	1498	1488	1457	1509	1391
V	0	3540	4075	4163	4844	4534	4461	4889	4964	4758

Table 11: Analysis of amino acid occurrence of complete human proteome.

Amino acid occurrence at first ten N-terminal positions of all human proteins.

Turnover Proteins (starting with M)										
Amino Acid	1st Position	2nd Position	3rd Position	4th Position	5th Position	6th Position	7th Position	8th Position	9th Position	10th Position
A	0	1866	908	706	658	623	607	597	656	603
R	0	234	447	475	481	494	519	492	489	513
N	0	172	177	196	160	166	199	180	157	154
D	0	292	373	217	259	296	245	219	267	265
C	0	46	94	104	85	93	114	105	117	115
Q	0	141	236	291	235	281	267	253	269	262
E	0	529	491	426	417	426	401	433	392	378
G	0	441	509	457	658	538	466	514	544	538
H	0	46	109	98	99	99	120	116	111	103
I	0	92	111	213	206	181	184	204	217	230
L	0	297	513	592	612	627	729	706	762	739
K	0	200	334	333	407	360	395	358	333	352
M	6402	101	98	121	128	113	95	103	91	102
F	0	125	145	238	177	182	204	194	216	178
P	0	375	383	491	357	416	455	470	414	480
S	0	823	639	567	560	560	540	497	487	485
T	0	261	357	298	290	294	313	309	281	269
W	0	84	82	92	79	94	68	95	77	98
Y	0	50	73	133	94	121	112	124	112	118
V	0	227	323	354	440	438	369	433	410	420

Table 12: Analysis of amino acid occurrence of turnover proteins.

Amino acid occurrence at the first ten N-terminal positions of proteins found in this study.

Top 10% Fastest Turnover Proteins (starting with M)										
Amino Acid	1st Position	2nd Position	3rd Position	4th Position	5th Position	6th Position	7th Position	8th Position	9th Position	10th Position
A	0	140	67	68	52	42	46	39	57	58
R	0	35	39	47	53	51	54	42	52	49
N	0	15	16	17	18	16	18	13	9	13
D	0	23	33	20	19	23	17	22	23	16
C	0	7	10	5	8	11	12	10	9	5
Q	0	9	29	36	19	28	30	22	28	19
E	0	49	24	25	38	37	40	51	29	37
G	0	48	36	41	54	37	41	38	44	46
H	0	2	8	6	6	6	7	9	9	10
I	0	10	11	17	15	21	15	18	16	24
L	0	21	51	37	51	59	58	61	68	65
K	0	18	27	26	29	32	30	25	35	33
M	558	8	7	15	15	15	8	15	3	16
F	0	13	10	24	16	10	15	20	17	17
P	0	29	42	47	32	45	42	41	42	35
S	0	73	59	47	49	56	45	44	46	49
T	0	25	36	26	34	30	24	28	16	16
W	0	7	6	9	7	10	11	7	9	8
Y	0	2	5	11	7	14	12	13	7	7
V	0	24	42	34	36	15	33	40	39	35

Table 13: Analysis of amino acid occurrence of fastest turnover proteins.

Amino acid occurrence for the top 10% fastest turnover proteins in this study.

Top 10% Slowest Turnover Proteins (starting with M)										
Amino Acid	1st Position	2nd Position	3rd Position	4th Position	5th Position	6th Position	7th Position	8th Position	9th Position	10th Position
A	0	156	88	69	66	57	56	48	57	59
R	0	24	41	36	31	44	44	41	45	40
N	0	9	10	12	13	14	15	19	13	8
D	0	13	21	12	22	24	17	14	26	26
C	0	5	7	10	7	8	11	12	13	10
Q	0	11	16	27	18	24	16	18	17	20
E	0	32	51	30	37	29	41	30	38	30
G	0	34	50	47	47	42	36	41	41	58
H	0	4	12	11	10	11	13	6	6	10
I	0	10	10	25	14	14	20	21	22	16
L	0	42	46	53	64	61	49	71	65	76
K	0	21	28	23	30	30	32	30	25	31
M	536	9	8	12	6	9	8	10	8	8
F	0	15	20	16	12	22	11	20	18	13
P	0	38	26	37	30	37	46	44	31	33
S	0	66	36	40	45	41	56	28	34	39
T	0	18	30	28	27	22	24	22	21	18
W	0	4	11	5	4	1	3	9	4	9
Y	0	8	6	12	8	3	8	10	13	6
V	0	17	19	31	45	43	30	42	39	26

Table 14: Analysis of amino acid occurrence of slowest turnover proteins.

Amino acid occurrence for the top 10% slowest turnover proteins in this study.

E. LabTracker: iPad Based Laboratory Management Software

The LabTracker project was an extension to the main PepTracker work discussed in this thesis. An honours project student, Yasir Ahmad - whom I supervised, carried out this project. The software was then integrated into PepTracker and further developed by myself.

The LabTracker software aims to provide an online electronic laboratory book and associated tools for researchers working at a laboratory bench. The need for this software arose as a direct result of the common problems encountered with traditional paper based laboratory books. Traditional laboratory books suffer from many issues including:

- Inefficient recording of laboratory activities,
- Keeping the laboratory book up-to-date,
- Keeping track of reagents,
- Lack of ability to search for data,
- Data can become easily unorganised,
- Legibility of written laboratory books, and
- Potential of loss and/or damage to the laboratory book.

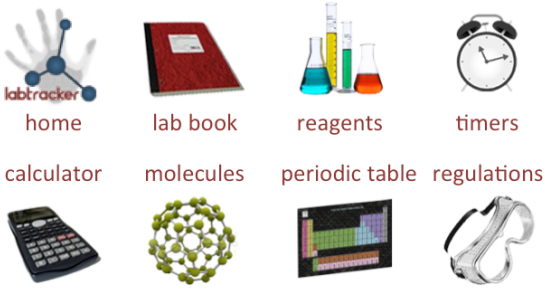
All of these issues highlight the need to move to an electronic based system that can record the everyday work carried out by a researcher, as well as provide added security that the data is backed up and protected. Furthermore, having this data in an electronic format means it can be easily searched and organised, is available from any computer connected to the web and it is legible and structured. Furthermore, by recording protocols and experiments on a web server means duplication can be avoided as protocols can be easily shared and kept consistent.

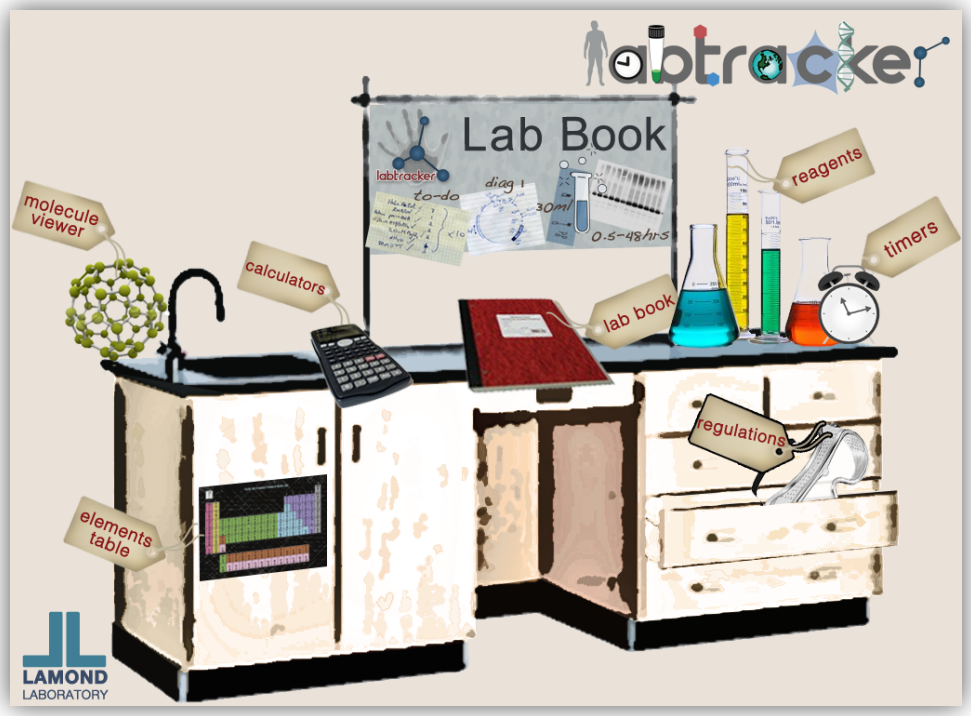
In order to implement a portable electronic notebook, an appropriate device had to be selected. For this work, the iPad was chosen as the platform of choice due to its

popularity in the scientific community and its ease of use. The iPad device is intuitive to use, lightweight and very portable, making an ideal option as a replacement for a traditional laboratory book in a science environment. Hence, the LabTracker software was developed as an Apple app that runs on iPad/iPhone devices.

It was decided that user laboratory data collected via the application should be stored on a central database server from where it is backed up. Each researcher can use an iPad with the LabTracker software installed to access their laboratory book on the server. This also provides the option of extending the application in the future to a web-based platform that can be accessed from any web browser on any device.

Shown below is the LabTracker homepage, which has links to all of the main sections in LabTracker: Lab Book, Reagents Database, Timers, Calculators, Molecule Viewer, Periodic Table and Health & Safety Regulations.





‘Lab Book’

The main component within LabTracker is the ‘Lab Book’. The ‘Lab Book’ provides researchers with a diary interface, allowing them to record the experiments they carry out on a daily basis, maintain a task list for each day, upload images from the in-built camera on the iPad 2 and setup alerts. When recording experiments being carried out, a researcher has the option of selecting from a library of standard protocols. This library is editable so that new protocols can be added and existing protocols updated. This ensures that protocols can be easily managed and maintained.

As a researcher carries out the steps within a protocol, they have the option of marking the individual steps as being complete, as well as adding notes to each step for future reference. With the iPad 2 allowing photo acquisitions, researchers can also take photographs of gels etc., at their laboratory bench, and have these stored with protocols.

Reagents Database

LabTracker also implements access to the reagents database held by PepTracker. LabTracker allows researchers to search for reagents from the reagent database. This is a particularly useful feature for researchers working at a laboratory bench. In the future this functionality could be extended to link up with the University ordering system to allow reagents to be ordered via the LabTracker software.

Timers

One of the useful tools for a researcher working on experiments at the laboratory bench is their timer. The LabTracker software implements the possibility of setting up electronic timers. Researchers can use this functionality to time different experimental stages, furthermore this functionality can be used to setup timer alerts for meetings etc. Researchers have the option of either visual and/or sound alerts.

Calculators

LabTracker also includes a series of scientific calculators. These include molarity and DNA/RNA conversion calculators. Without this functionality being at hand at the laboratory bench, researchers would have to return to their office computer to carry out these calculations.

Molecule Viewer

The 3D molecule viewer provides researchers with an interactive interface to view different molecule structures. Currently this includes all amino acid structures but could be extended in the future to other structures, for example protein structures. Researchers have the option of rotating molecules in the x, y and z planes, alternating between ball and stick and letter representations, as well as free rotation with hand gestures.

Periodic Table

An interactive periodic table of elements is implemented, providing easy to access information on all elements in a pop-up style window.

Health and Safety Regulations

It is imperative health and safety regulations are followed within a laboratory setting. In order to make this easier for researcher, health and safety documentation is made available via the LabTracker application for easy access. This information includes guidance on radioactive substances, risk assessments, fume hood operation etc.

The LabTracker software is under continued development and evaluation by researchers in the Lamond Laboratory. The tools and feature sets available via the application are being updated and reviewed to ensure they are effective in their purpose and usable. It is intended that this software could be extended to other platforms in the future via a web interface that communicates data from the central LabTracker database.